Zuur Ch 04 Umm… heterogeneity

HARUG! QRantine edition

Ed Harris

Heterogeneity

Y = fixed part + random part

$\alpha + \beta_1 X_1 + \cdots + \beta_q X_q$

$\alpha + f_1(X_1) + \cdots + f_q(X_q)$

Heterogeneity

Nested data (random effects)

Temporal correlation

Spatial correlation

Random noise

Models we know and love

The *vanilla* **linear regression** model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where

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

Models we know and love

The **non-linear regression** (GAM) model

$$Y_i = \alpha + f(X_i) + \epsilon_i$$

where

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

RE terminology (Relatable)

The confusing aspects of most of these books are the wide range of different names and underlying mathematical notation. Mixed modelling, multilevel analysis, hierarchical linear models, and repeated measurements are just a few of the names that all refer to the same set of models.

General problem here

How to deal with unequal variation

- Transformation is traditional
- Modelling should fit the problem… - The variation is often *important*

This chapter is about “weighted regression” (like we saw in the Dunn and Smyth)

Loligo forbesi (long-finned squid)

Dataset *Squid.rdata*

```
##   Specimen YEAR MONTH DML Testisweight
## 1     1017 1991     2 136         0.006
## 2     1034 1990     9 144         0.008
## 3     1070 1990    12 108         0.008
## 4     1070 1990    11 130         0.011
## 5     1019 1990     8 121         0.012
## 6     1002 1990    10 117         0.012
```

Dataset *Squid.rdata*

```
table(Squid$Specimen)[1:5]
##
## 1001 1002 1003 1004 1005
##   50   43   36   32   29
```

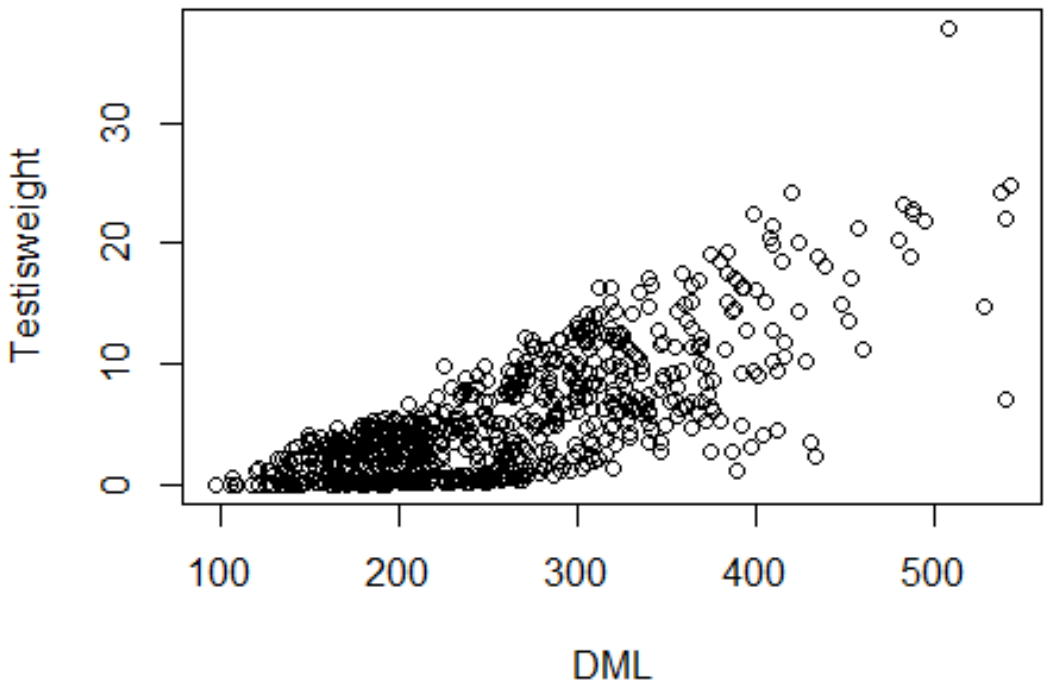
Dataset *Squid.rdata*

Factors explaining variation in sexual maturity (measured by testis size)

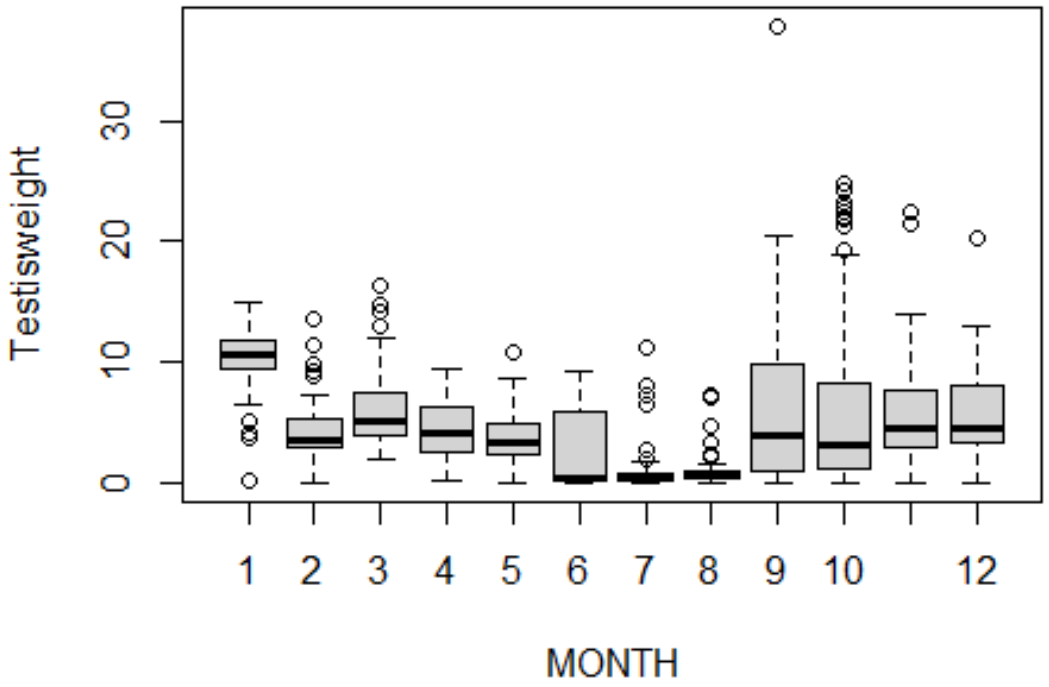
Testisweight ~ DML * Month + ϵ_i (eq 4.1)

Interaction term; Month is considered *nominal*; homogeneity of variance assumption

Dataset *Squid.rdata* NB bad aesthetics



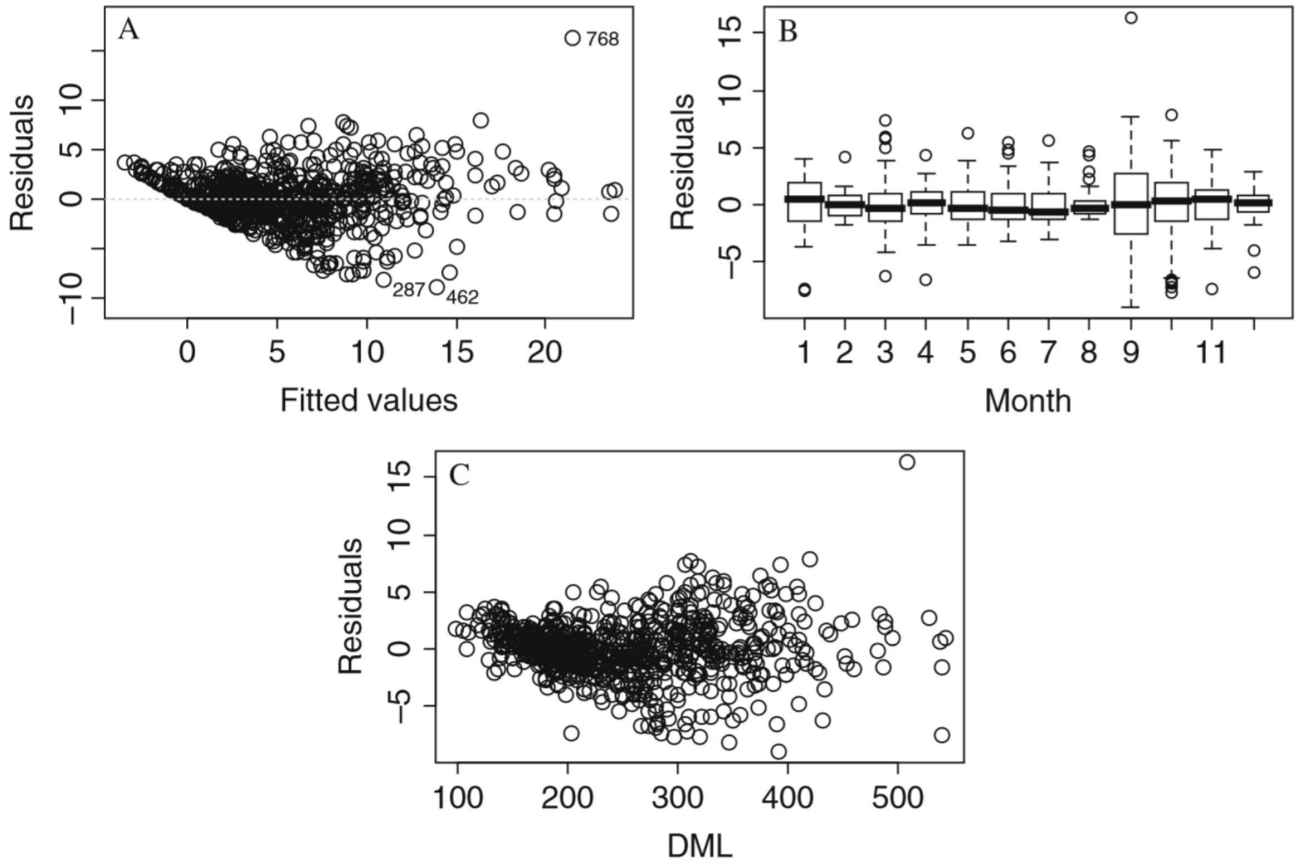
Dataset *Squid.rdata*



Dataset *Squid.rdata*

```
Squid$fMONTH <- factor(Squid$MONTH)

M1 <- lm(Testisweight ~ DML * fMONTH, data =
Squid)
```

Dataset *Squid.rdata*



Dataset *Squid.rdata*

Assumption $\epsilon_i \sim N(0, \sigma^2)$

Residual error mean of 0, homogeneous variance

Clearly wrong!

Looks like variance increases with body size…

Option #1 set a “fixed variance”

Zuur suggests explicitly accounting for the association between variance and body size

The approach here is exactly the same as weighted regression - set the error variance to scale with body size (by multiplying it by body size…)

where $\epsilon_i \sim N(0, \sigma^2 \times DML_i)$

Dataset *Squid.rdata*

```
# NB na.exclude
SquidNNA <- na.exclude(Squid)


# == regular linear model
M.lm <- gls(Testisweight ~ DML * fMONTH, data
= SquidNNA)


# accounts for increase in variance ~DML
M.gls1 <- gls(Testisweight ~ DML * fMONTH,
weights = formula(~DML), data = SquidNNA)
```

Dataset *Squid.rdata*

```
anova(M.lm, M.gls1)
##         Model df      AIC      BIC    logLik
## M.lm        1 25 3752.084 3867.385 -1851.042
## M.gls1      2 25 3620.898 3736.199 -1785.449
```

Dataset *Squid.rdata*

```
Anova(M.gls1)
## Analysis of Deviance Table (Type II tests)
##
## Response: Testisweight
##             Df   Chisq Pr(>Chisq)
## DML          1 1396.39  < 2.2e-16 ***
## fMONTH      11  321.17  < 2.2e-16 ***
## DML:fMONTH 11  201.97  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1
```

Option #2 “VarIdent”

Recognise variance change due to MONTH

$$Testisweight_{ij} = intercept + DML_{ij} + Month_j + DML_{ij}{:}\,Month_j + residuals_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma_j^2), where\ j = 1{:}\,12(Months)$$

Option #2 “VarIdent”

```
vf2 <- varIdent(form = ~ 1 | fMONTH)

M.gls2 <- gls(Testisweight ~ DML*fMONTH, data =
SquidNNA, weights = vf2)
```

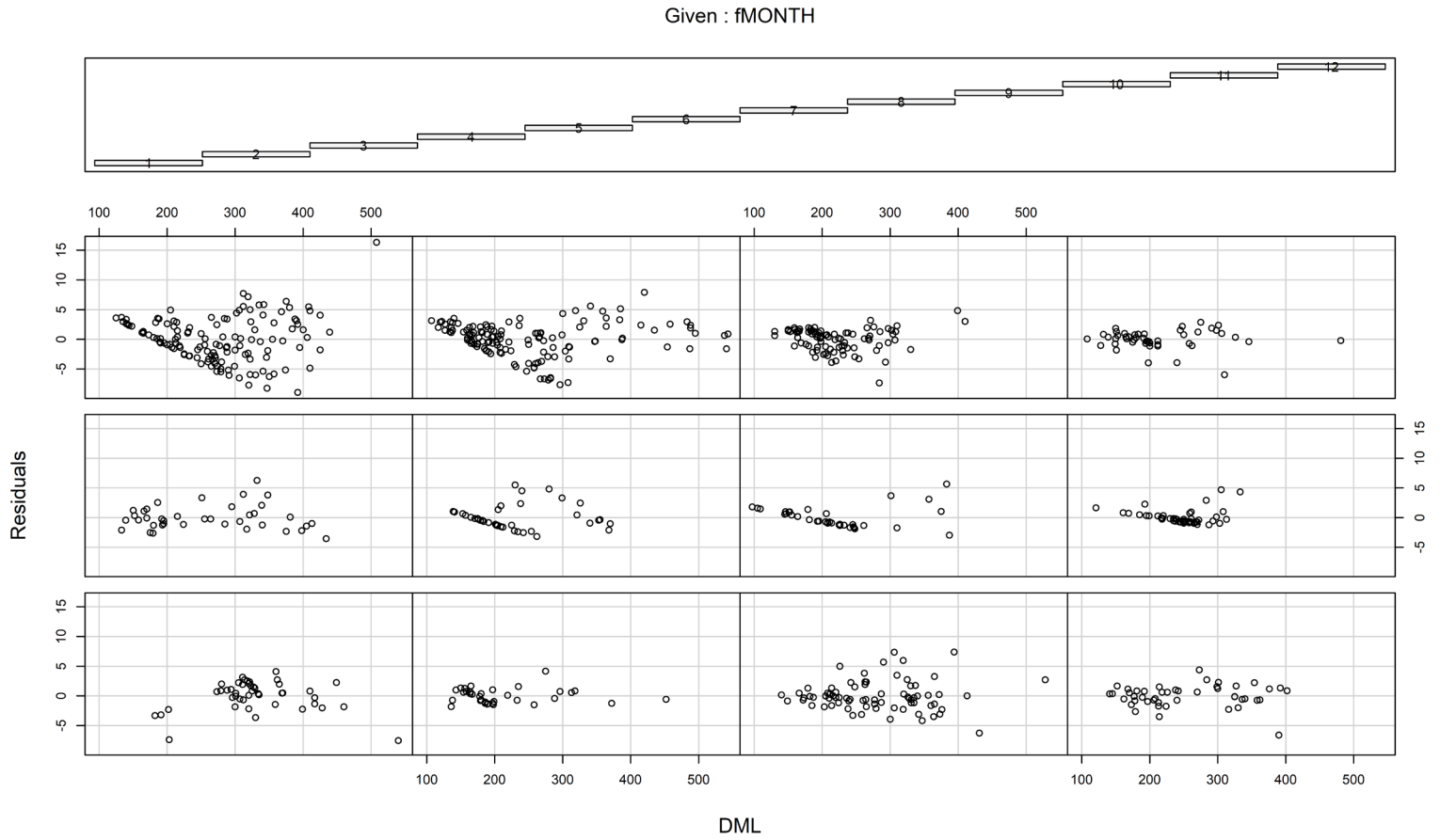
Option #2 “VarIdent”

```
# NB comparing model with and without
# month error, but apples to oranges
# to compare models with body size versus
# month errors structure
vf2 <- varIdent(form = ~ 1 | fMONTH)

M.gls2 <- gls(Testisweight ~ DML*fMONTH, data =
SquidNNA, weights = vf2)

anova(M.lm, M.gls2)

##         Model df      AIC      BIC    logLik
Test  L.Ratio p-value

## M.lm       1 25 3752.084 3867.385 -1851.042

## M.gls2     2 36 3614.436 3780.469 -1771.218
1 vs 2 159.6479  <.0001
```

Option #2 “VarIdent”

```
> summary(M.gls2)
...
Variance function:
Structure: Different standard deviations per stratum
Formula: ~1 | fMONTH
Parameter estimates:
2    9    12   11   8    10   5    7    6    4
1.00 2.99 1.27 1.50 0.98 2.21 1.63 1.37 1.64 1.42
1    3
1.95 1.97
...
Residual standard error: 1.27
```

Option #2 “VarIdent”



Option #2 “VarIdent”

- Fixes issue for some months
- Some months still have variance issues
- Uneven sample sizes
- Probably need to account for both…
- Message: take responsibility for your own error structure! (via trial and error, stats awareness, subject knowledge)

Option #3 varPower

Raise the weighting factor of error variance by some exponent

$$\epsilon_i j \sim N(0, \sigma^2 \times |DML_{ij}|^{2\delta}$$

where δ is an estimated value…

Option #3 varPower

```
vf3 <- varPower(form = ~DML)
vf4 <- varPower(form = ~DML | fMONTH)

M.gls3 <- gls(Testisweight ~ DML * fMONTH,
weights = vf3, data = SquidNNA)
M.gls4 <- gls(Testisweight ~ DML * fMONTH,
weights = vf4, data = SquidNNA)
```

Option #3 varPower

$\varepsilon_{ij} \sim N(0, \sigma^2 \times |DML_{ij}|^{2\delta}$ (eq. 4.6)

Accounts for both DML and MONTH

Option #3 varPower

Dank note: small typo between eq. 4.5 & 4.6

```
## [1] 3407.511
## Analysis of Deviance Table (Type II tests)
##
## Response: Testisweight
##             Df  Chisq Pr(>Chisq)
## DML          1 819.76  < 2.2e-16 ***
## fMONTH      11 781.94  < 2.2e-16 ***
## DML:fMONTH 11 263.77  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question

The ANOVA tables for these are all similar, all main effects are significant, etc.

Why bother with all this fiff faff with the residuals?

Small digression

We have seen:
varFixed()
(weighted error by some [continuous, numeric] vector)

varIdent()
(weighted error by some factor)

varPower() (weighted error by some power function)

NB on AIC

We are using “model selection” (AIC) a lot

It is basically essential to be aware of this

NB on AIC

Ecology, 95(3), 2014, pp. 631–636
© 2014 by the Ecological Society of America

Model selection for ecologists: the worldviews of AIC and BIC

Ken Aho,[1,4] DeWayne Derryberry,[2] and Teri Peterson[3]

Option #4 varExp

This is technically complicated beyond the Zuur book

Basically you can incorporate an exponential weighted structure to the variance of the residuals

We will quickly look at this

Option #4 varExp

$\varepsilon_{ij} \sim N(0, \sigma^2 \times e^{2\delta \times DML_i})$ (eq. 4.7)

Option #4 varExp

```
vf5 <- varExp(form = ~DML|fMONTH)

M.gls5 <- gls(Testisweight ~ DML * fMONTH,
weights = vf5, data = SquidNNA)
```

Option #4 varExp

```
AIC(M.gls5)
## [1] 3419.719
Anova(M.gls5)
## Analysis of Deviance Table (Type II tests)
##
## Response: Testisweight
##             Df  Chisq Pr(>Chisq)
## DML          1 829.97  < 2.2e-16 ***
## fMONTH      11 799.14  < 2.2e-16 ***
## DML:fMONTH 11 162.99  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*'
0.05 '.' 0.1 ' ' 1
```

Option #5 varConstPower

Basically you can incorporate a **constant** and an **exponential** weighted structure to the variance of the residuals

Variants for DML and DML + MONTH

Option #5 varConstPower

$\varepsilon_{ij} \sim N(0, \sigma^2 \times (\delta_1 + |DML_{ij}|^{\delta_2})^2$ (eq. 4.8)

Option #5 varConstPower

```
vf6 <- varConstPower(form = ~DML)
vf7 <- varConstPower(form = ~DML | fMONTH)

M.gls6 <- gls(Testisweight ~ DML*fMONTH,
weights = vf6, data = SquidNNA)
M.gls7 <- gls(Testisweight ~ DML*fMONTH,
weights = vf7, data = SquidNNA)
```

Option #5 varConstPower

```
AIC(M.gls4, M.gls5, M.gls6, M.gls7)
##         df      AIC
## M.gls4 37 3407.511
## M.gls5 37 3419.719
## M.gls6 27 3475.019
## M.gls7 49 3431.511
```

Option #6ish varComb

You can use varComb to mix and match other variance structures.

These models were not better than Mgls.4, but it is an option that might make practical sense…

```
vf8 <- varComb(varIdent(form = ~1 | fMONTH), varExp(form = ~DML) )
```

Pinheiro and Bates 2000, pp.214

Table 4.1 Various variance structures used in this section. The table follows Pinheiro and Bates (2000)

Name of the function in R	What does it do?
VarFixed	Fixed variance
VarIdent	Different variances per stratum
VarPower	Power of the variance covariate
VarExp	Exponential of the variance covariate
VarConstPower	Constant plus power of the variance covariate
VarComb	A combination of variance functions

Caveats and deciding which is best

Caveats and deciding which is best

- Trial and error
- Responsibility for investigating assumptions
- Model selection with caveats
- Subject specific knowledge of your own variance issues

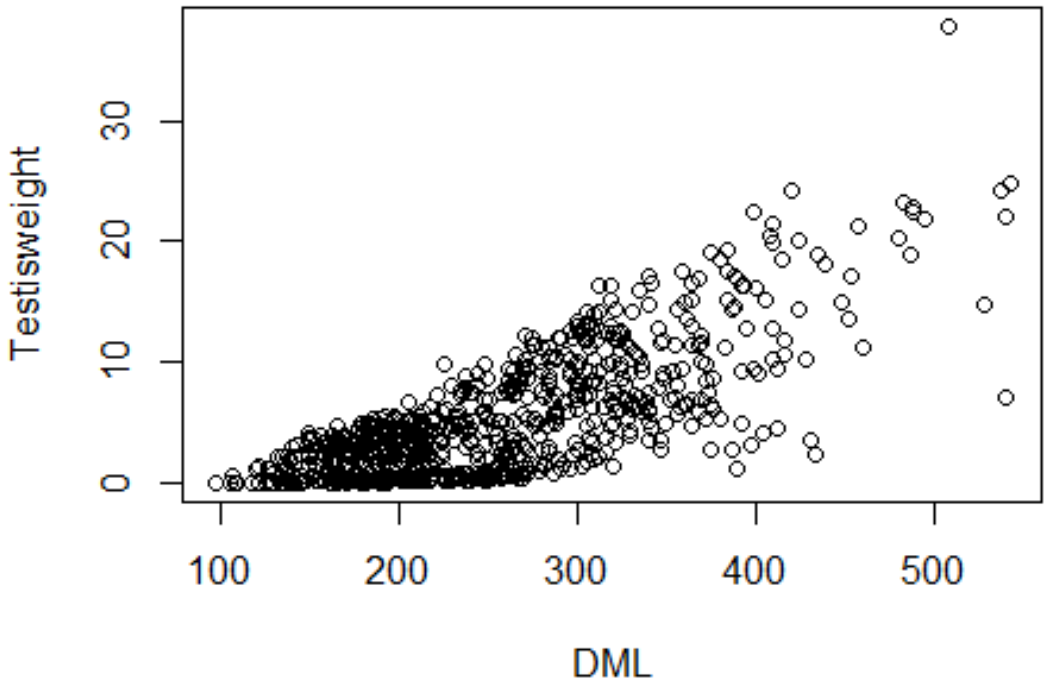
Caveats and deciding which is best

Graphical evaluation of residuals

ordinary residuals
(obs - fitted values)

standardised residuals
(difference relative to variance structure)

Simple(r) example than Zuur



Simple(r) example than Zuur

```
mygls1 <- gls(Testisweight ~ DML, data =
SquidNNA)


myvf <- varPower(form = ~DML)
mygls2 <- gls(Testisweight ~ DML,
              weights = myvf, data = SquidNNA)
```

Simple(r) example than Zuur

```
anova(mygls1, mygls2)
```

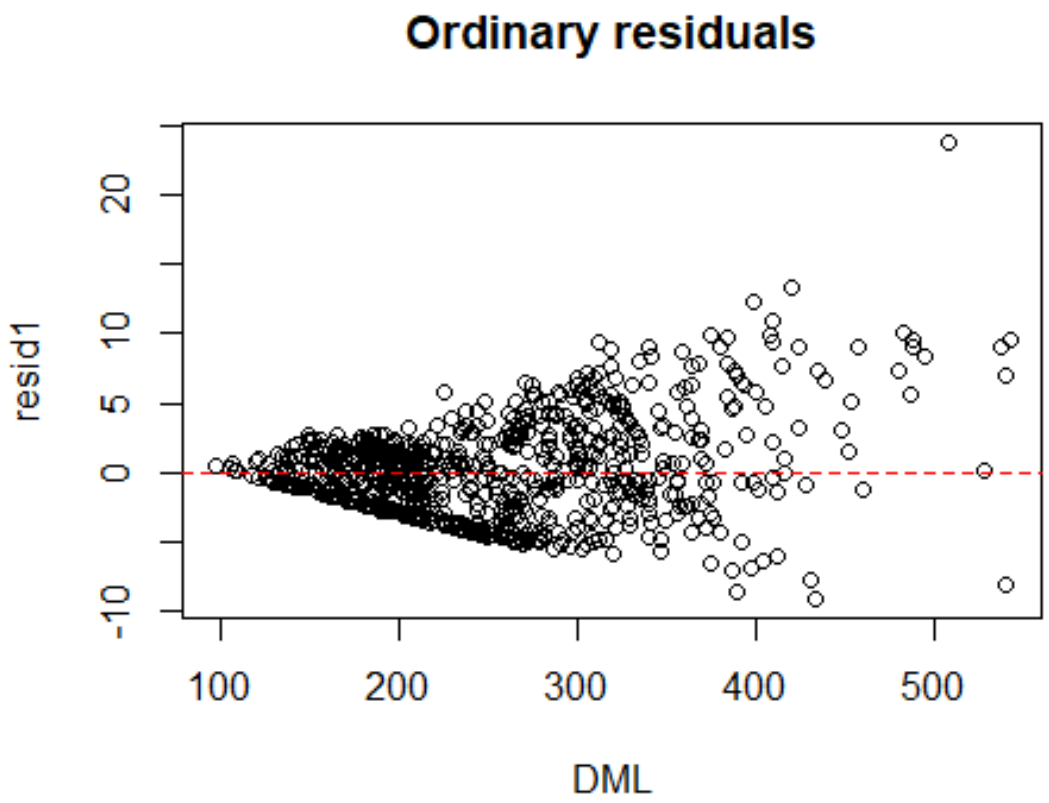
```
##        Model df      AIC      BIC    logLik
Test  L.Ratio p-value
## mygls1     1  3 4055.094 4069.018 -2024.547
## mygls2     2  4 3725.415 3743.980 -1858.707
1 vs 2 331.6796  <.0001
```

Simple(r) example than Zuur

```
Anova(mygls2)
## Analysis of Deviance Table (Type II tests)
##
## Response: Testisweight
##     Df  Chisq Pr(>Chisq)
## DML  1 720.91  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

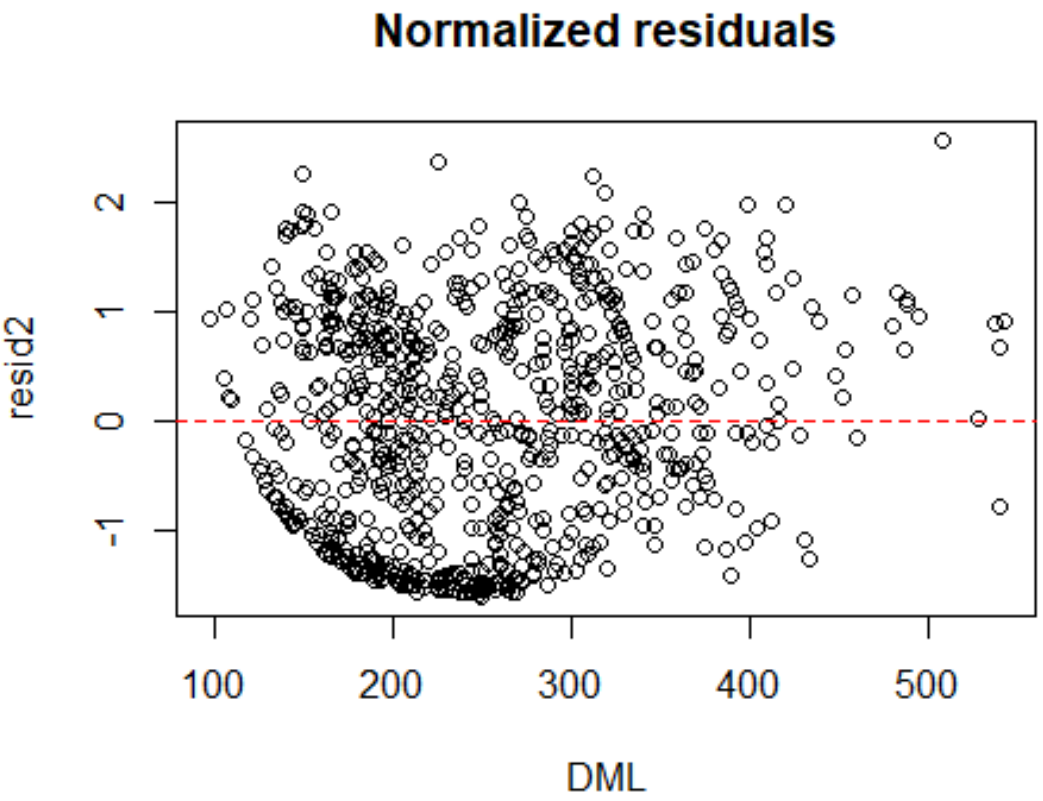
Simple(r) example than Zuur

```
resid1 <- resid(mygls2)
plot(resid1 ~ DML, data = SquidNNA,
     main = "Ordinary residuals")
abline(h=0, col="red", lty=2)
```


Ordinary residuals
resid1
20
10
5
0
-10
100
200
300
400
500
DML

Simple(r) example than Zuur

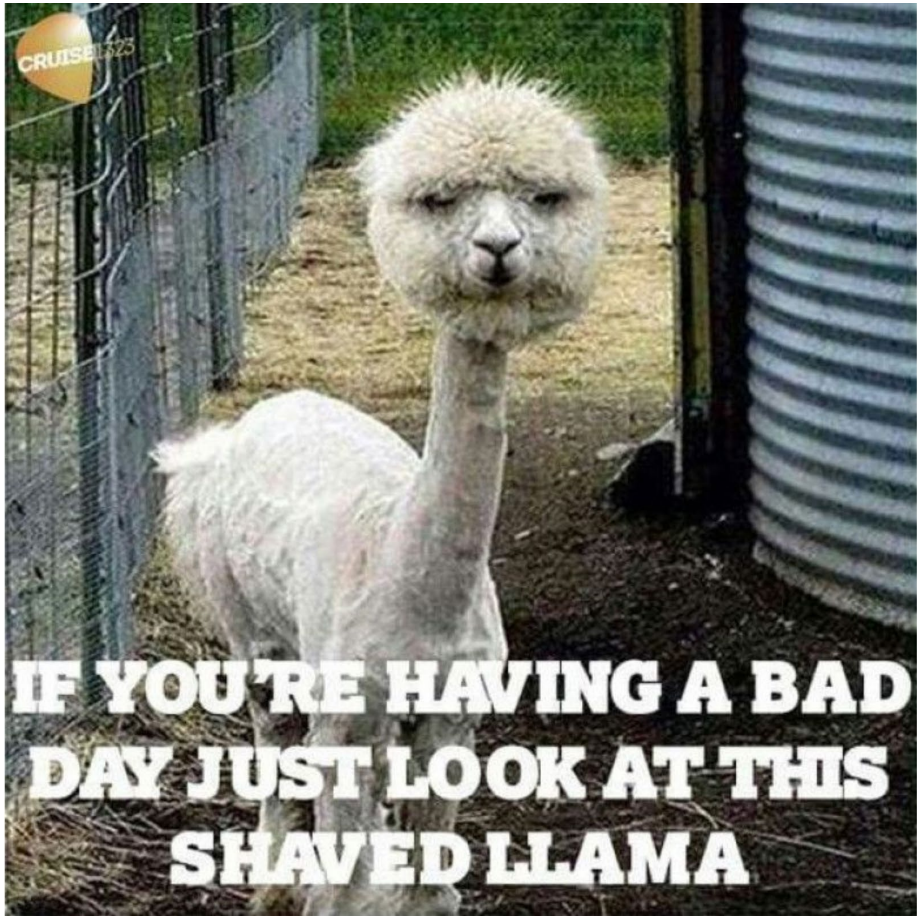
```
resid2 <- resid(mygls2, type = "normalized")
plot(resid2 ~ DML, data = SquidNNA,
     main = "Normalized residuals")
abline(h=0, col="red", lty=2)
```


Normalized residuals
resid2
-1
0
1
2
100
200
300
400
500
DML

Simple(r) example than Zuur

How do you think ordinary versus normalized residuals will compare for a model with the (typical case) assumption of fixed residual variance?

You look like you have had a bad data day

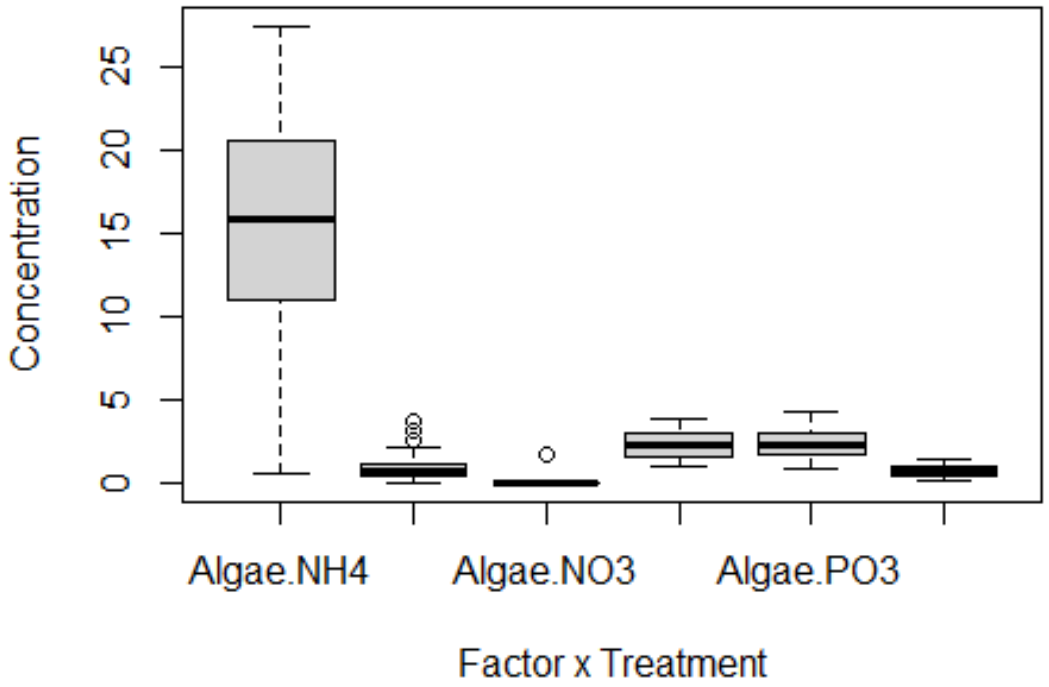


Benthic Biodiversity data

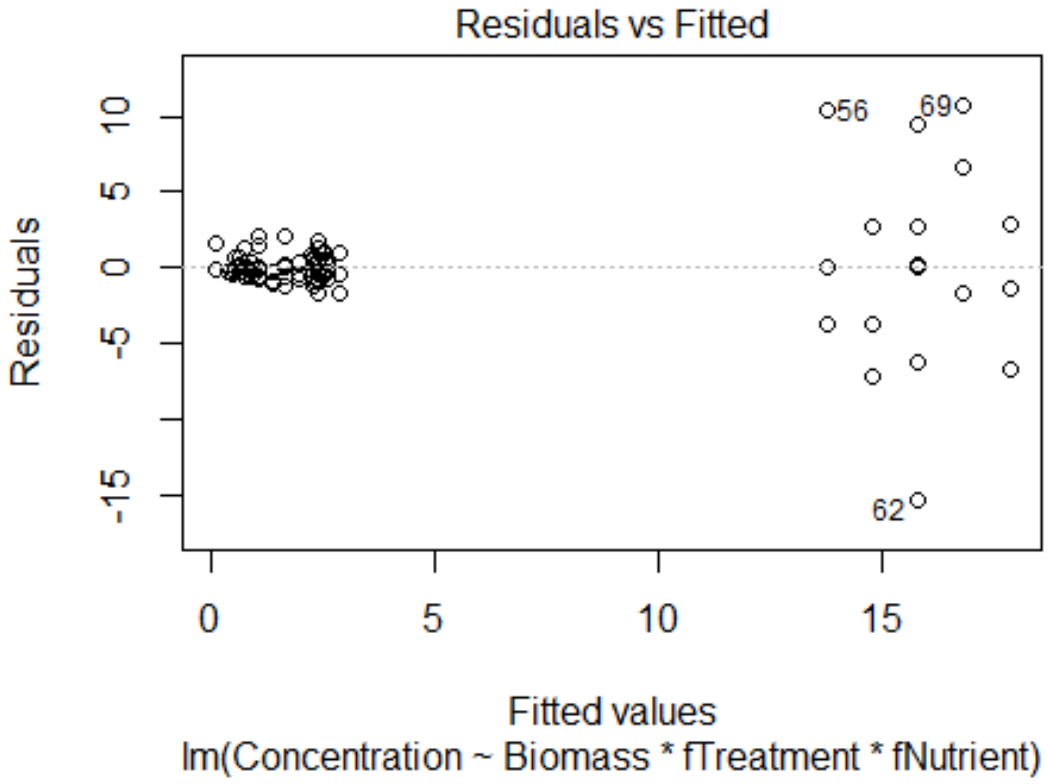
Benthic Biodiversity data

	Abundance	Treatment	Nutrient	Concentration
1	0	NoAlgae	NO3	2.490
2	0	NoAlgae	NO3	1.185
3	0	NoAlgae	NO3	3.825
4	4	NoAlgae	NO3	3.045
5	4	NoAlgae	NO3	2.190
6	4	NoAlgae	NO3	3.600

Benthic Biodiversity data



Benthic Biodiversity data



Benthic Biodiversity data

```
f1 <- formula(Concentration ~ Biomass *
fTreatment * fNutrient)

M0 <- gls(f1, data = Biodiv)

M1A <-gls(f1, data = Biodiv, weights =
varIdent( form = ~1 | fTreatment * fNutrient))

M1B <-gls(f1, data = Biodiv, weights =
varIdent(form = ~1 | fNutrient))

M1C <-gls(f1, data = Biodiv, weights =
varIdent(form = ~1 | fTreatment))
```

Benthic Biodiversity data

```
anova(M0, M1A, M1B, M1C)
```

```
##      Model df      AIC      BIC    logLik   Test   L.Ratio p-value
## M0       1 13 534.5203 567.8569 -254.2602
## M1A      2 18 330.1298 376.2881 -147.0649 1 vs 2 214.39054  <.0001
## M1B      3 15 380.0830 418.5482 -175.0415 2 vs 3  55.95320  <.0001
## M1C      4 14 439.7639 475.6647 -205.8819 3 vs 4  61.68087  <.0001
```

Benthic Biodiversity data

```
Analysis of Deviance Table (Type II tests)

Response: Concentration
                           Df    Chisq Pr(>Chisq)
Biomass                     1   1.2218   0.269010
fTreatment                  1   6.6649   0.009833 **
fNutrient                   2   3.1551   0.206483
Biomass:fTreatment          1   1.9062   0.167387
Biomass:fNutrient           2   8.3560   0.015329 *
fTreatment:fNutrient        2 243.1430  < 2.2e-16 ***
Biomass:fTreatment:fNutrient  2   2.1809   0.336071
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The Protocol (for model selection)

1.1 Start with full model (explanatory variables + interactions)

1.2 Explore assumption of homoegeneity of variance of residual error

1.3 I would personally insert some simple model transformation here…

The Protocol (for model selection)

2.1 Repeat step 1 using gls() from {nlme}

(the reason for this is to set up nested model comparisons tweaking error structure)

The Protocol (for model selection)

3.1 Compare competing models using specific error structure (NB this requires knowledge and experience, but even if you have neither you are respoonsible for your own assumptions)

3.2 Compare resulting models and residual error structure

The Protocol (for model selection)

4.1 Fit fresh gls() using specific error structure you have dicided on

(specify argument method=REML)

The Protocol (for model selection)

5.1 Model comparison using AIC, etc.

5.2 Final test of assumpotions for best or "close@ models

5.3 NB that model selection < model average, the latter of which is “a thing” now

The Protocol (for model selection)

6.1 Consider alternative error distribution assumptions

(like the GLM Poisson, etc.)

6.2 Transformation here only as last resort (diverge slightly from my opinion)

The Protocol (for model selection)

7 Prepare models for comparison

7.1.a All possible models

7.1.b Only specific subset of possible models

7.2 If you get serious here, use liklihood ratio for ML method versions…

The Protocol (for model selection)

8 Perform model comparison

9 Validation and assumption scrutiny for REML version of best model

10 Prepare results and discuss meaning