General linear models (not Generalized linear model)

Linear Model	Common name
$Y = \mu + X$	Simple linear regression
$Y = \mu + A_1$	One-factorial (one-way) ANOVA
$Y = \mu + A_1 + A_2 + A_1 \times A_2$	Two-factorial (two-way) ANOVA
$Y = \mu + A_1 + X\ (+A_1 \times X)$	Analysis of Covariance (ANCOVA)
$Y = \mu + X_1 + X_2 + X_3$	Multiple regression
$Y = \mu + A_1 + g + A_1 \times g$	Mixed model ANOVA
$Y_1 + Y_2 = \mu + A_1 + A_2 + A_1 \times A_2$	Multivariate ANOVA (MANOVA)

Y (response) is a continuous variable

X (predictor) is a continuous variable

A represents categorical predictors (factors)

g represents groups of data (more on this later)

$(+A_1 \times X)$ - step 1 on an ANCOVA, but not in the final analysis

Multiple factors $A_1 + A_2 + \text{etc (and their interactions)}$

Understanding and dealing with heterogeneity

Intermediary steps before
going fully mixed…..

…….. model

Let's start with a problem

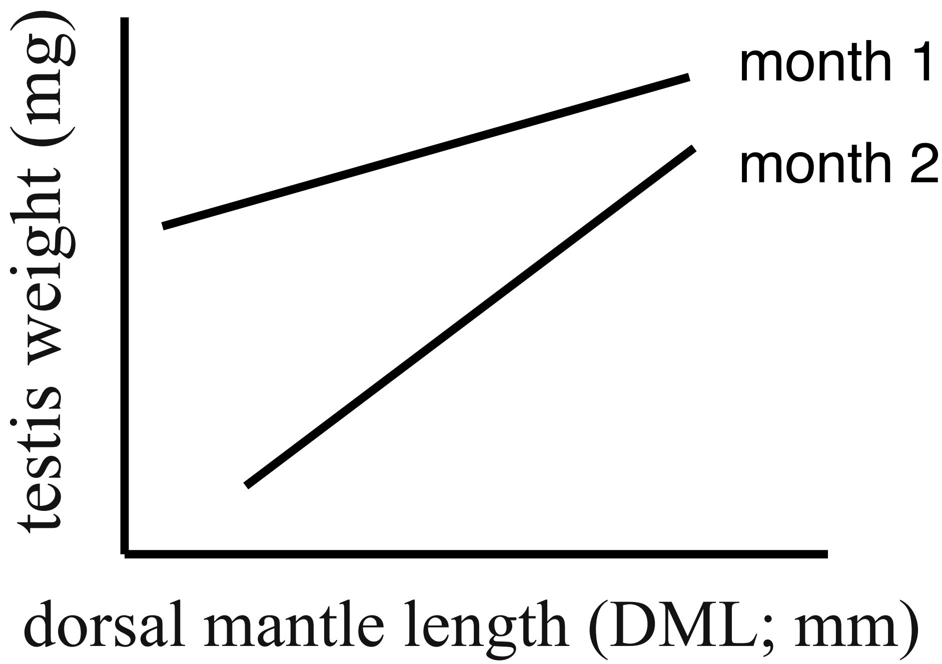
Seasonal patterns of investment in reproductive and somatic tissues in the squid *Loligo forbesi*

Jennifer M. Smith[1,a], Graham J. Pierce[1], Alain F. Zuur[2] and Peter R. Boyle[1]

[1] Department of Zoology, School of Biological Sciences, University of Aberdeen, Tillydrone Avenue, Aberdeen AB24 2TZ, UK
[2] Highland Statistics Ltd., 6 Laverock Road, Newburgh, Aberdeenshire, AB41 6FN, UK

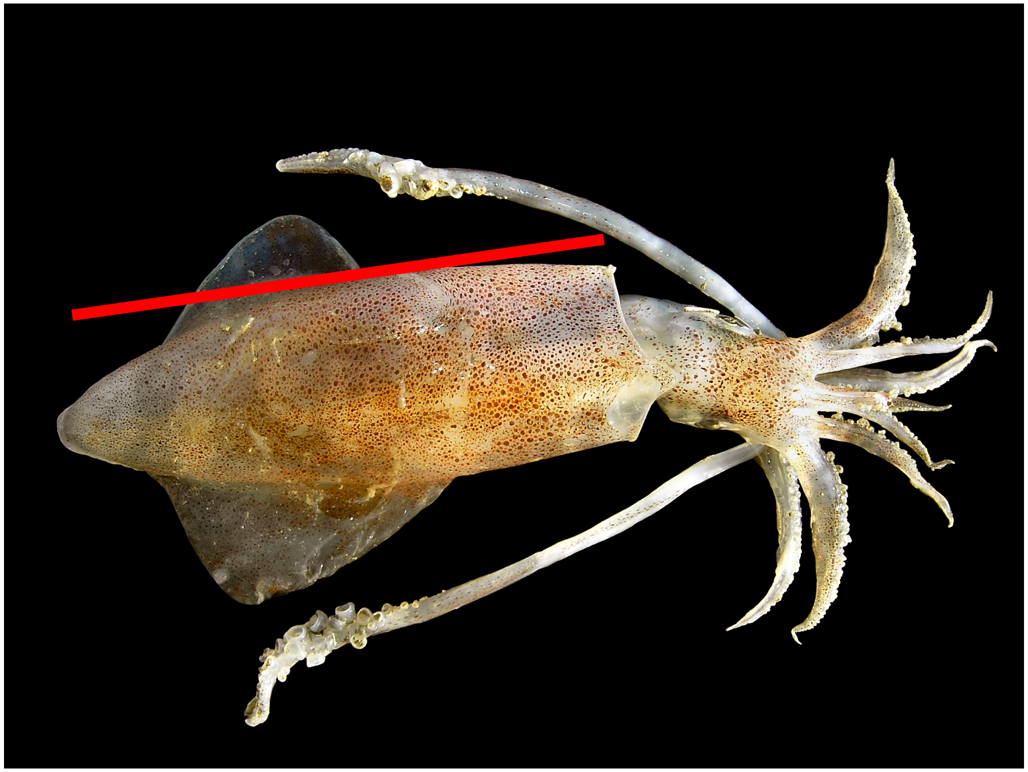
Goal: study seasonal variation (patterns) in reproductive and somatic tissues (mating is aseasonal).

In which month there is more investment (relative to individual size, i.e., DML) in reproduction?

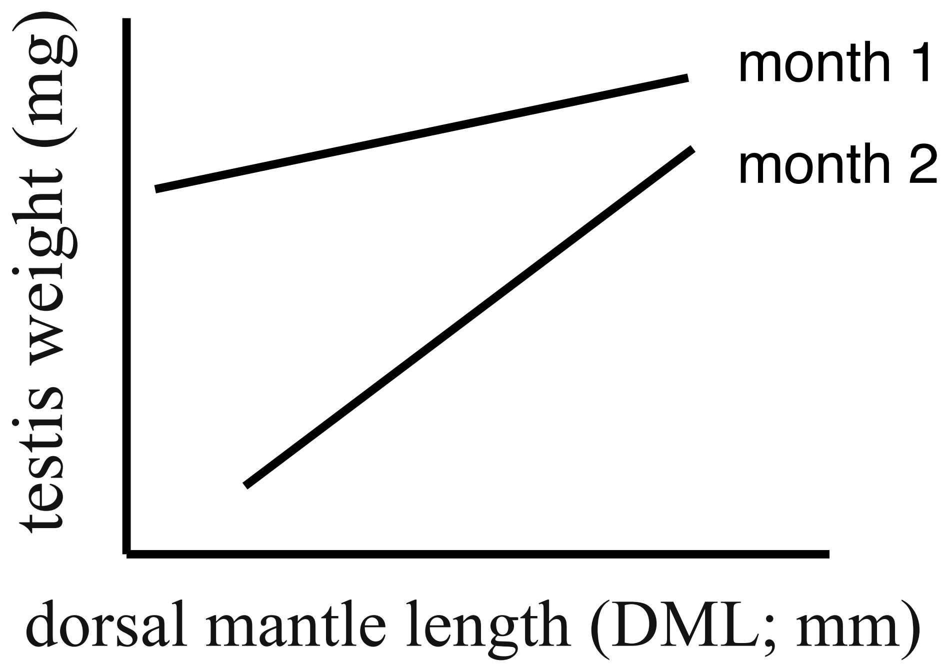


Aquat. Living Resour. 18, 341–351 (2005)

DOI: 10.1051/alr:2005038
www.edpsciences.org/alr



Goal: study seasonal patterns in reproductive and somatic tissues.



In which month there is more investment (relative to individual size DML) in reproduction?

Data structure

	A	B	C	D
1	Specimen	MONTH	DML	Testisweight
2	1017	2	136	0.006
3	1034	9	144	0.008
4	1070	12	108	0.008
5	1070	11	130	0.011
6	1019	8	121	0.012
7	1002	10	117	0.012
8	1001	5	133	0.013
9	1013	7	105	0.015
10	1002	7	109	0.017
11	1006	7	97	0.017
12	1020	9	144	0.022
13	1002	6	141	0.023
14	1039	9	125	0.024
15	1038	9	140	0.026
16	1012	12	128	0.027
17	1037	9	142	0.036
18	1001	6	139	0.036
19	1027	7	145	0.043
20	1003	7	181	0.05

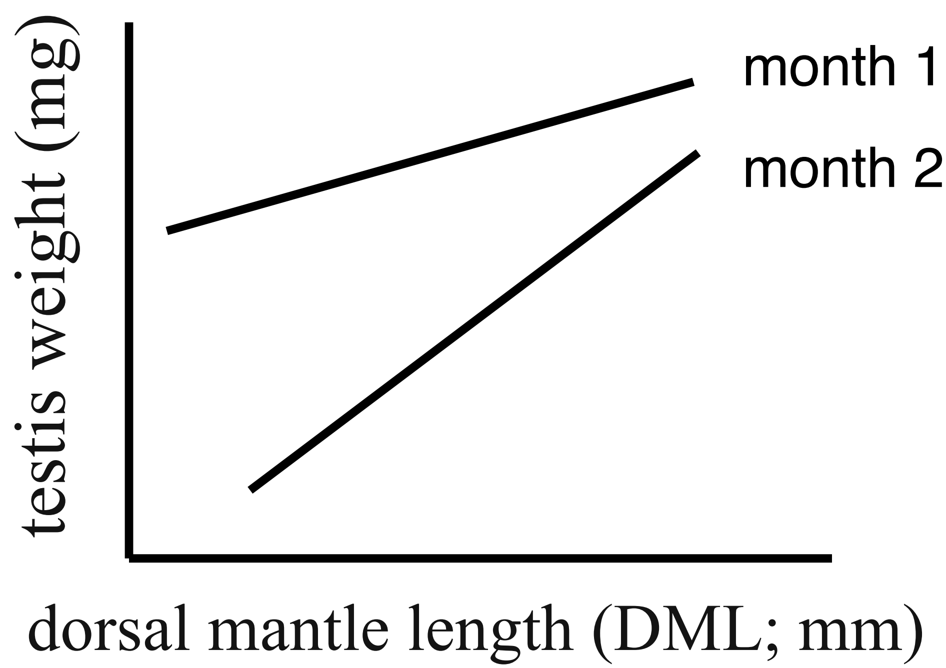
.

.

.

768 individuals

Goal: study seasonal patterns in reproductive and somatic tissues.



Goal: study seasonal patterns in reproductive and somatic tissues.

Model of interest

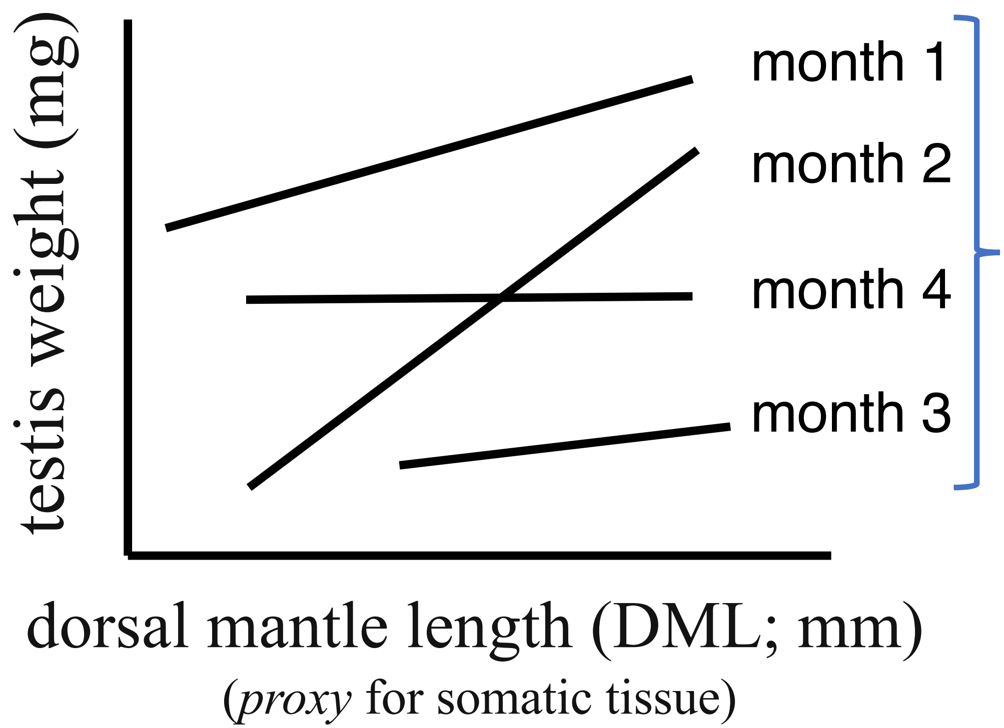
$$\text{TestisWeight} = \text{constant} + \beta_1\text{DML} + \beta_2\text{Month} + \beta_3(\text{DML} \times \text{Mont}) + \ \text{e}$$

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

continuous variable

continuous variable

categorical variable (factor)



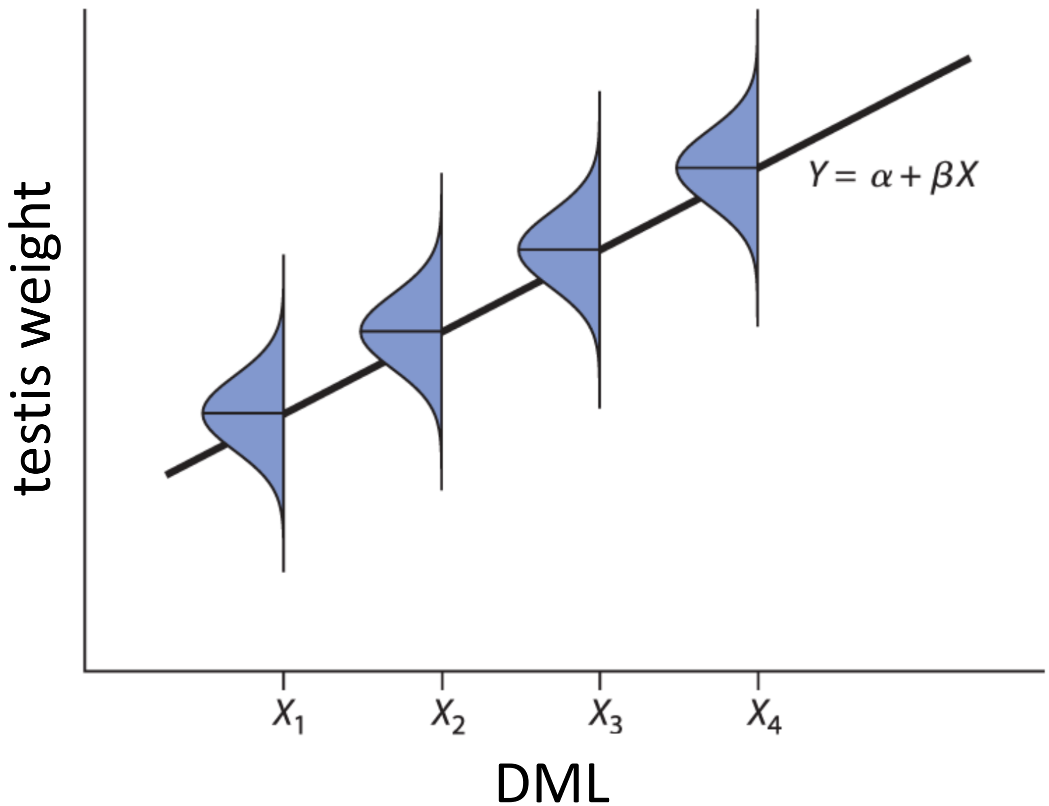
seasonal variation (environmental drivers)?

What component of the model quantifies and test for the variation in slopes across months?

déjà vu

The assumption of constant residual variance (homoscedasticity)

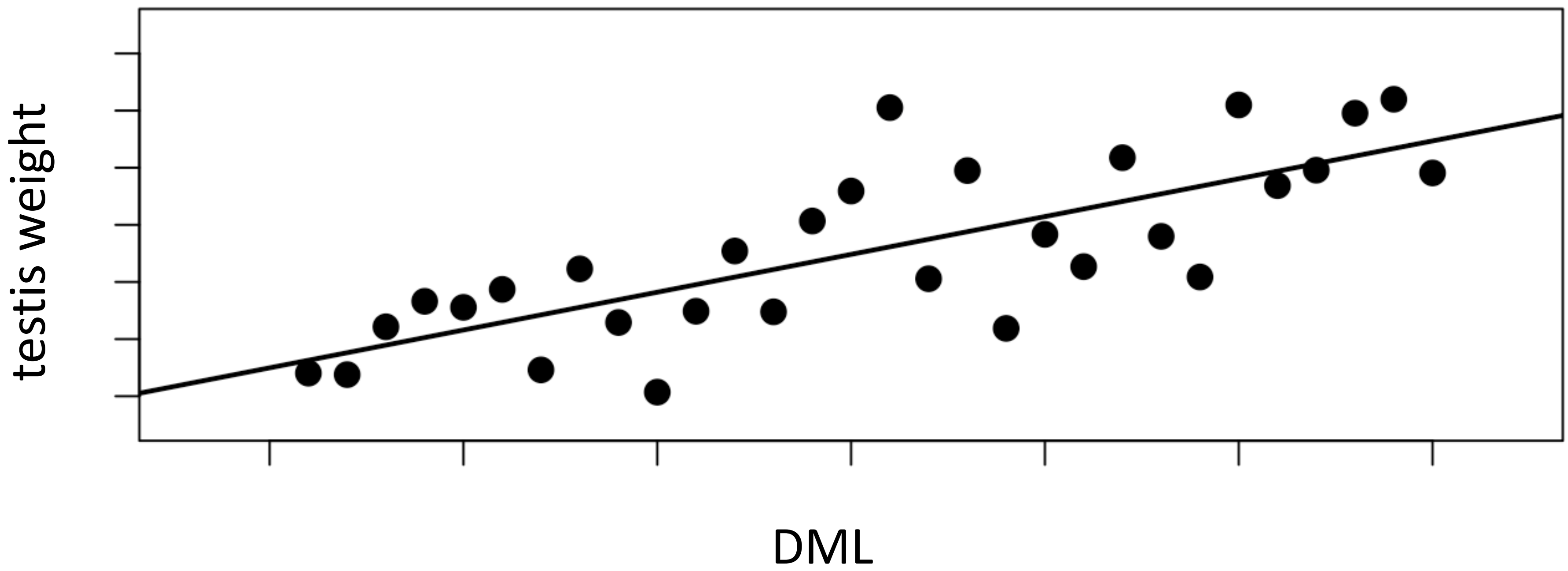
$$e \sim N(0, \sigma^2)$$



Understanding $e \sim N(0, \sigma^2)$

sampling variation in residual from the same population model

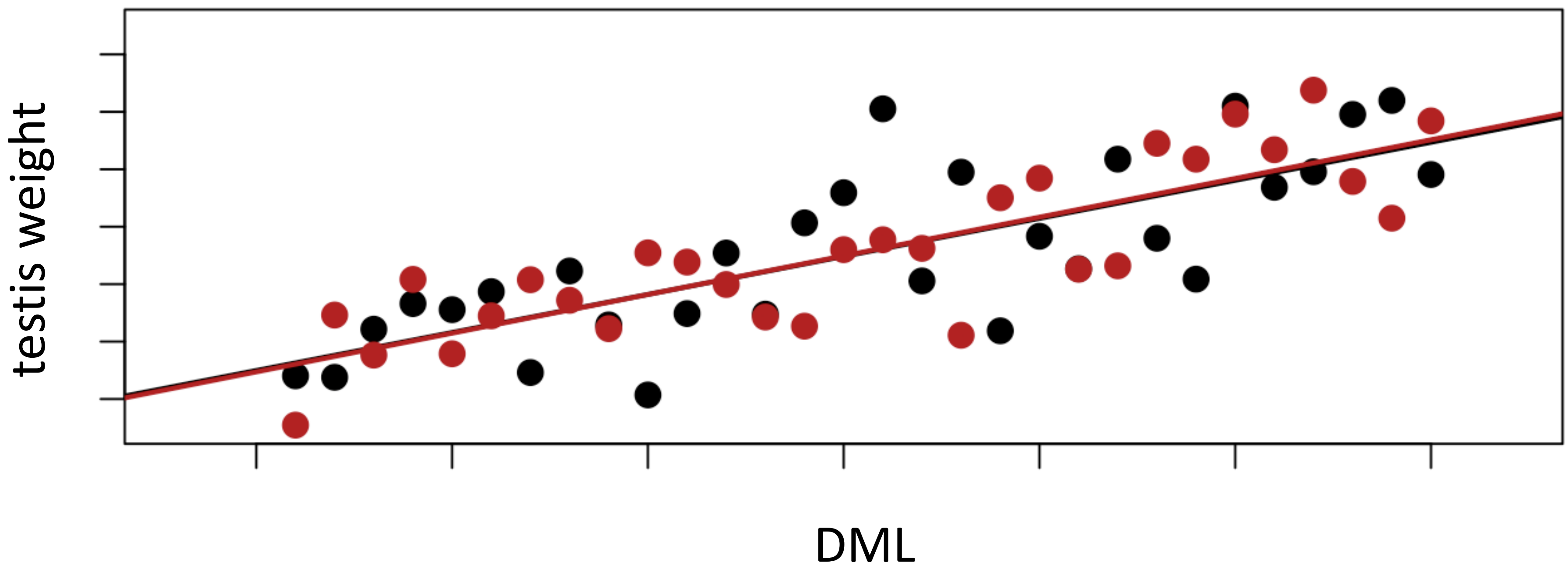
One possible sample



Understanding $e \sim N(0, \sigma^2)$

sampling variation in residual from the same population model

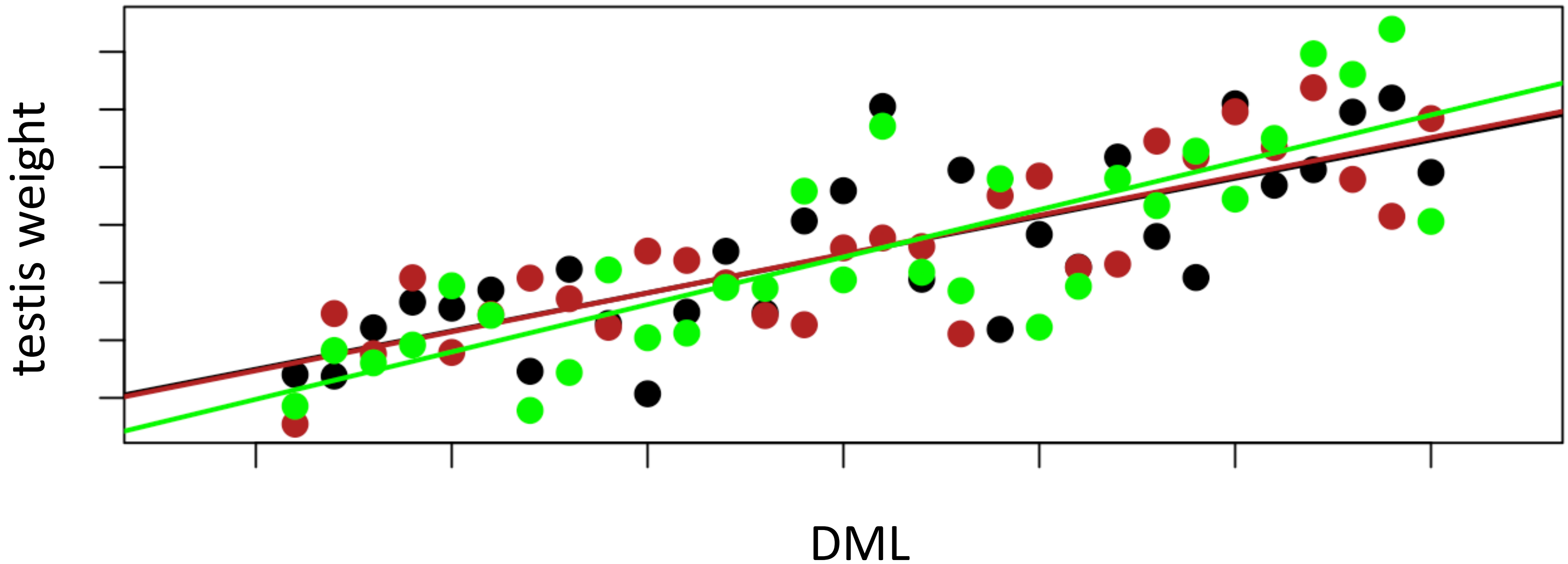
Another possible sample and the first possible sample



Understanding $e \sim N(0, \sigma^2)$

sampling variation in residual from the same population model

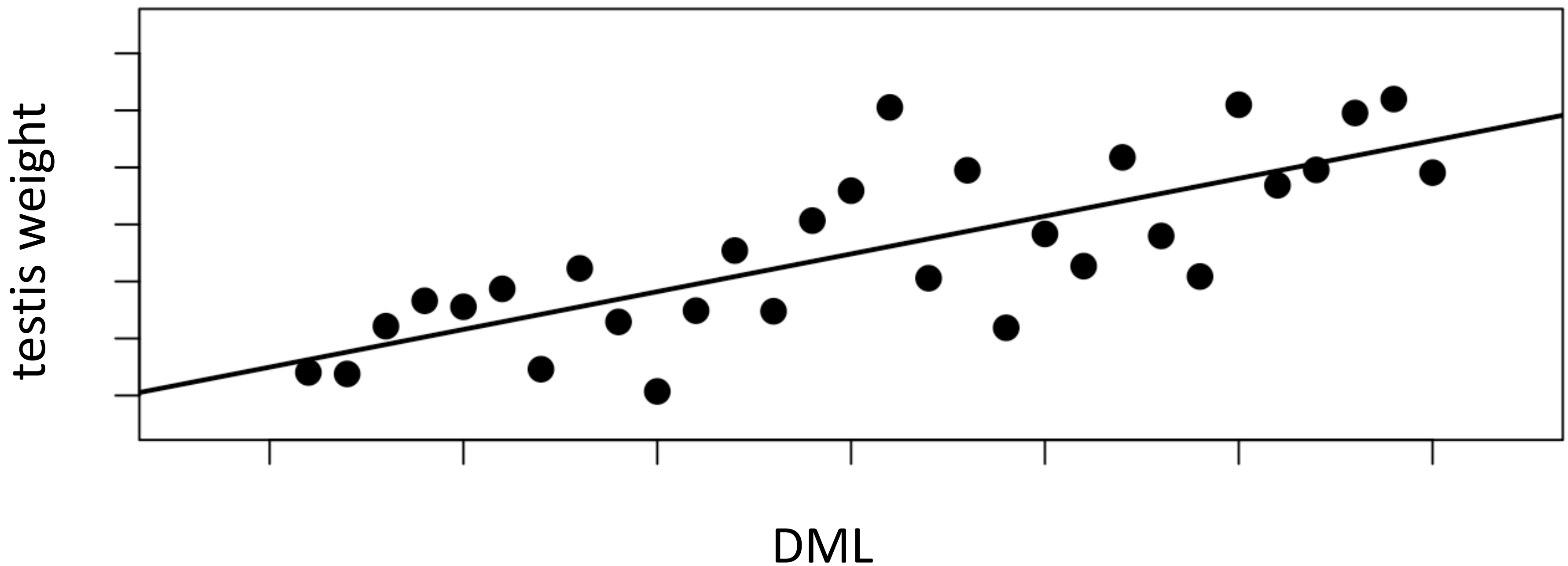
Yet another possible sample and the first two possible samples



Understanding $e \sim N(0, \sigma^2)$

sampling variation in residual from the same population model

One possible sample

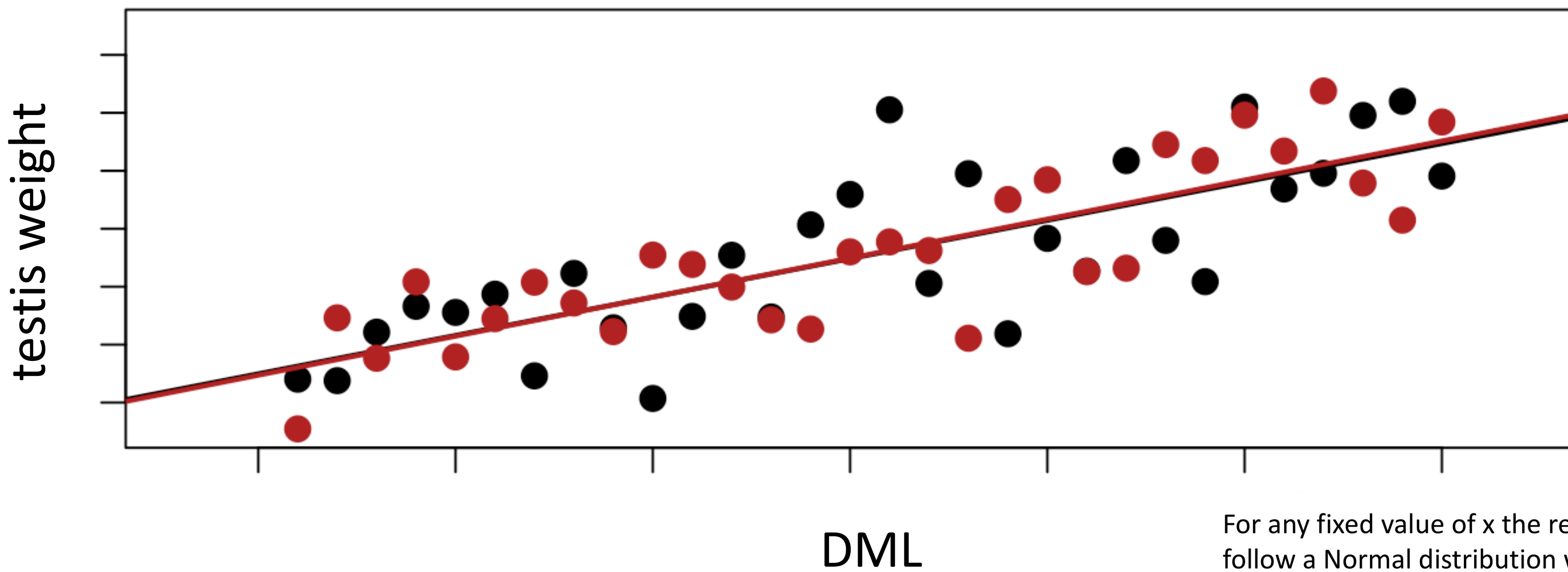


AGAIN

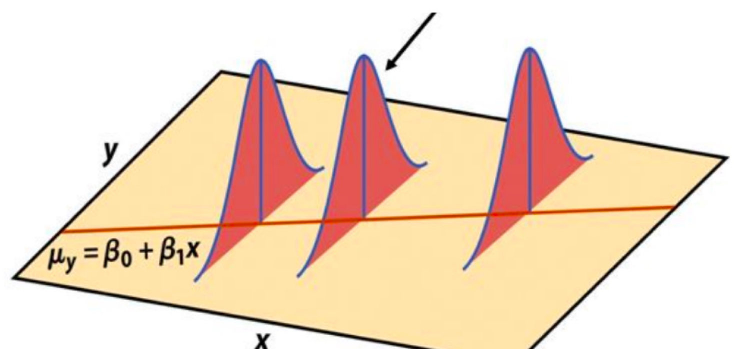
Understanding $e \sim N(0, \sigma^2)$

sampling variation in residual from the same population model

Another possible sample and the first possible sample



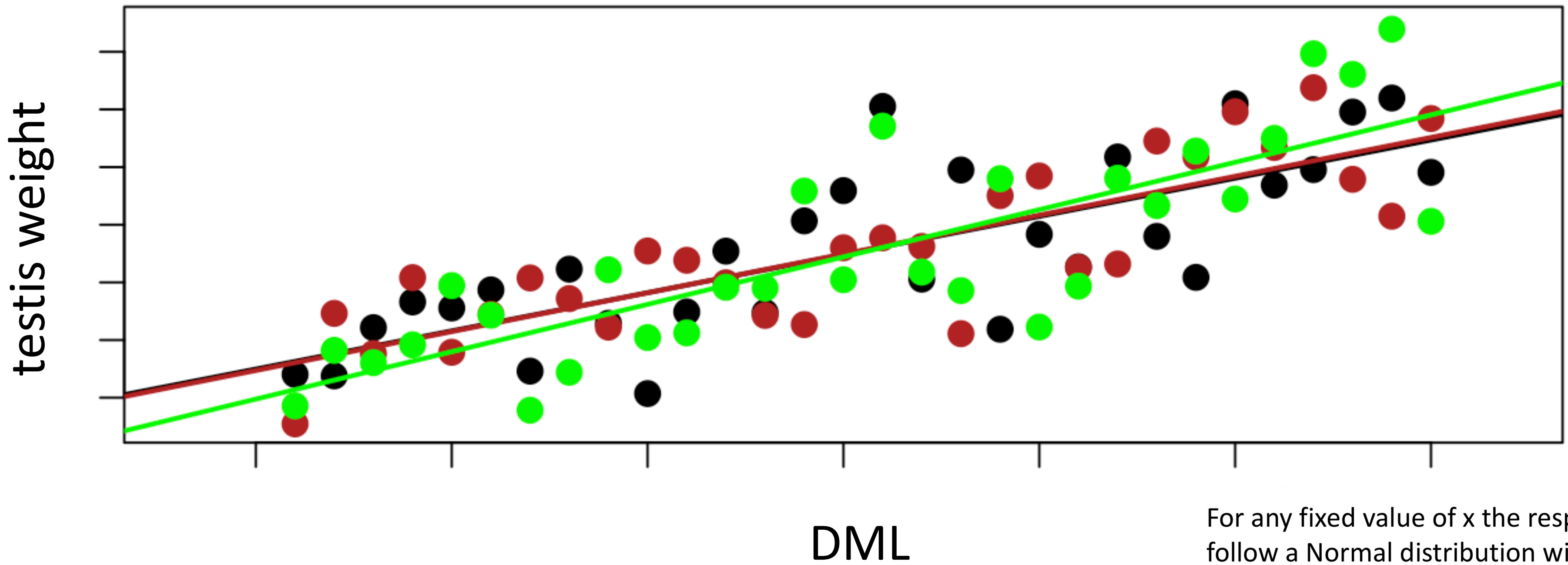
For any fixed value of x the responses y follow a Normal distribution with variance σ^2



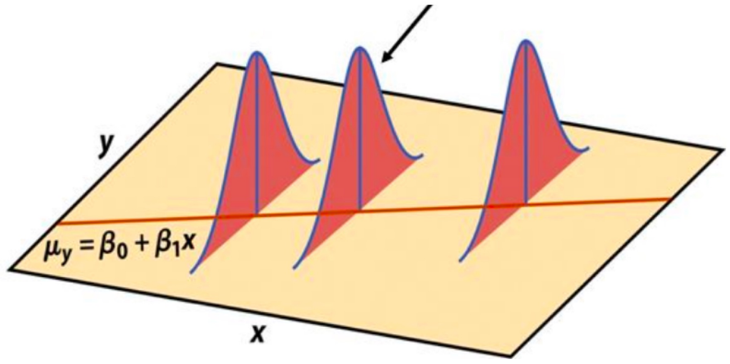
Understanding $e \sim N(0, \sigma^2)$

sampling variation in residual from the same population model

Yet another possible sample and the first two possible samples



For any fixed value of x the responses y follow a Normal distribution with variance σ^2



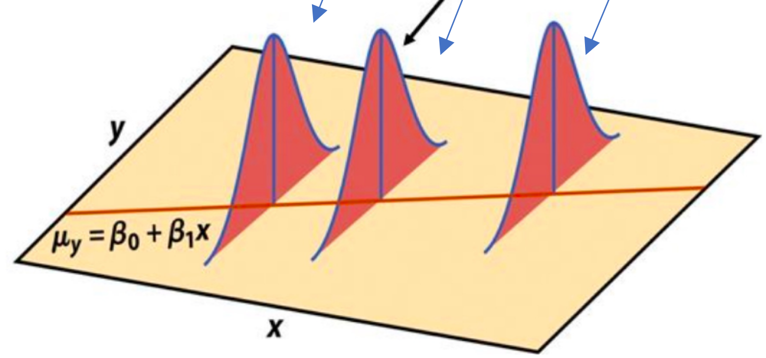
$e \sim N(0,\sigma^2) \sim N(0,\Sigma)$, i.e, HOMOscedasticity

homogeneity of variance (all variances in the diagonal are equal)

$$y_i = \underbrace{\beta_0 + \beta_1 \times x_i}_{\text{Linearity}} + \varepsilon_i \qquad \varepsilon_i \sim \underbrace{\mathcal{N}(0,\sigma^2)}_{\text{Normality}} \qquad \mathbf{V} = cov = \begin{array}{c} \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & \vdots \\ \vdots & \cdots & \sigma^2 & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{pmatrix} \text{Observations} \\ \text{Observation} \end{array}$$

Zero covariance (=independence)

For any fixed value of x the responses y follow a Normal distribution with variance σ^2



VARIANCES OF RESIDUALS ARE ASSUMED NOT TO VARY ACROSS OBSERVATIONS IN THE STANDARD REGRESSION MODEL

(called fixed variance structure)

VARIANCES OF RESIDUALS VARY ACROSS OBSERVATIONS IN THE MODEL (called variable [non-fixed] variance structure)

$e \neq N(0, \sigma^2) \neq N(0, \Sigma)$, i. e, HETEROscedasticity

Heteroscedasticity (variances in the diagonal are not equal)

$$y_i = \underbrace{\beta_0 + \beta_1 \times x_i}_{\text{Linearity}} + \varepsilon_i \qquad \varepsilon_i \sim \underbrace{\mathcal{N}(0, \sigma^2)}_{\text{Normality}} \qquad \mathbf{V} = cov = \begin{array}{cc} \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & \vdots \\ \vdots & \cdots & \sigma_3^2 & \vdots \\ 0 & \cdots & \cdots & \sigma_4^2 \end{pmatrix} & \text{Observations} \\ \text{Observation} & \end{array}$$

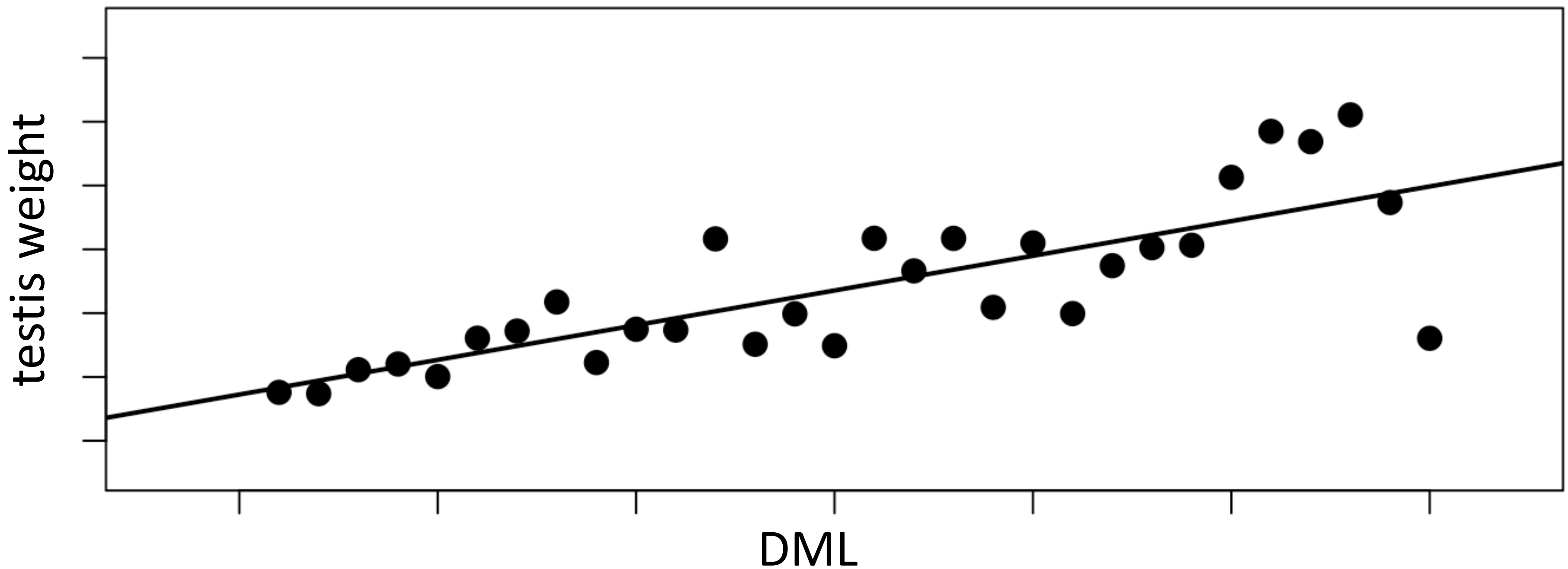
Zero covariance (=independence)

VARIANCES OF RESIDUALS VARY
ACROSS OBSERVATIONS IN THE MODEL
(called variable [non-fixed] variance structure)

$e \neq N(0,\sigma^2) \neq N(0,\Sigma)$, i.e, HETEROscedasticity

sampling variation in residual from the same population model

One possible sample

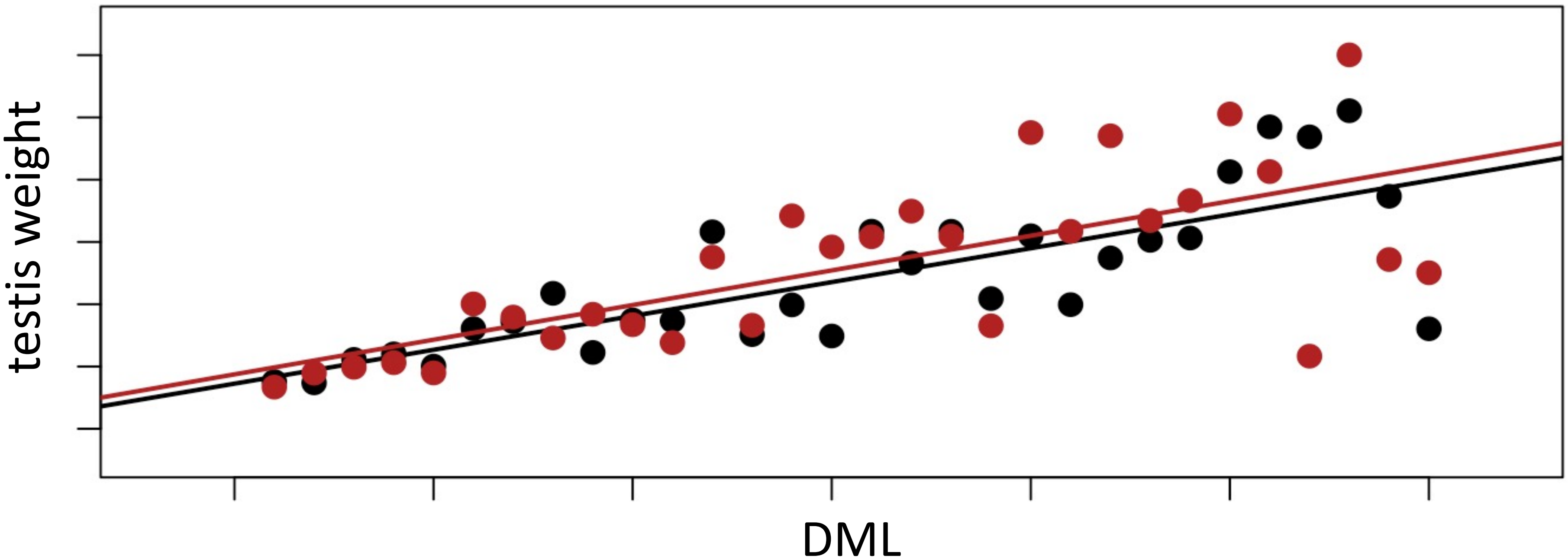


variance increasing with predictor (DML)

$e \neq N(0, \sigma^2) \neq N(0, \Sigma)$, i.e, HETEROscedasticity

sampling variation in residual from the same population model

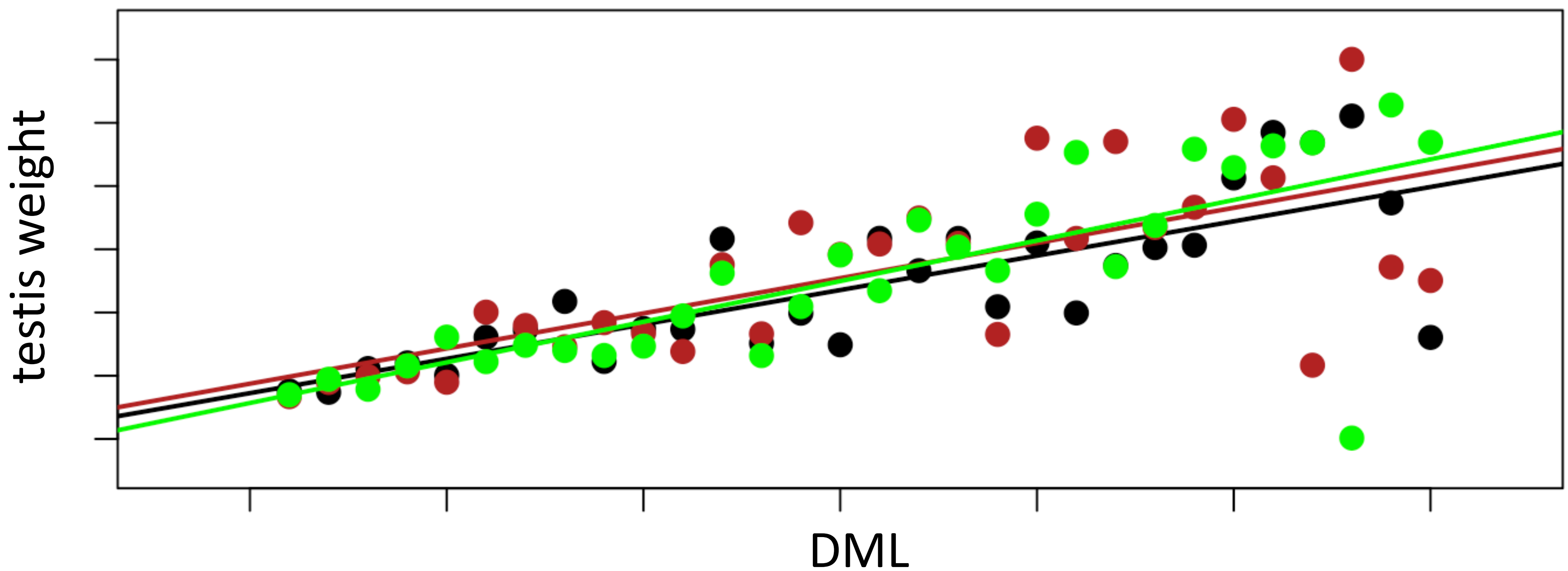
Another possible sample and the first possible sample



$e \neq N(0,\sigma^2) \neq N(0,\Sigma)$, i.e, HETEROscedasticity

sampling variation in residual from the same population model

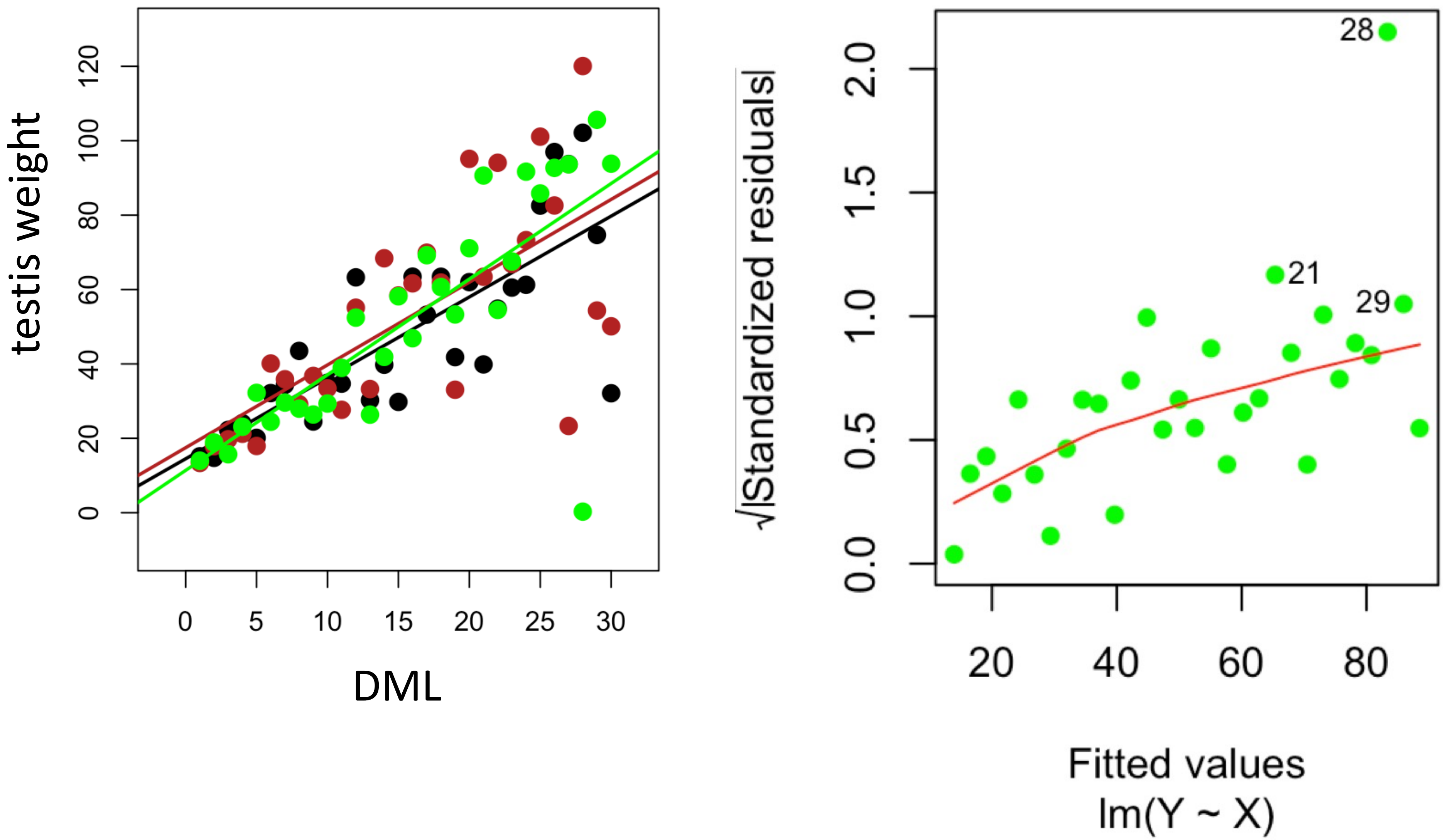
Yet another possible sample and the first two possible samples



$e \neq N(0, \sigma^2) \neq N(0, \Sigma)$, i. e, HETEROscedasticity

sampling variation in residual from the same population model

Yet another possible sample and the first two possible samples



$e \neq N(0,\sigma^2) \neq N(0,\Sigma)$, i.e, HETEROscedasticity

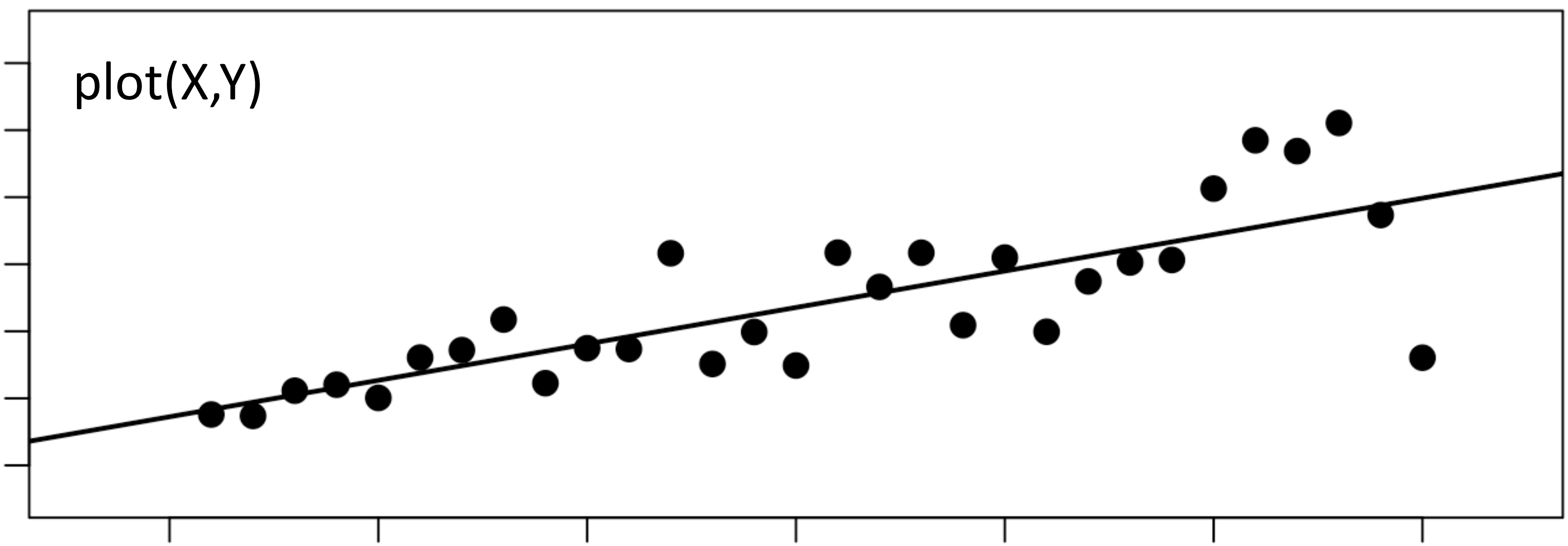
How was variance heterogeneity generated in these examples?

```
n=30
X = 1:n
e = rnorm(n,0,X)
Y = constant + slopeX * X + e
```

$e \neq N(0,\sigma^2) \neq N(0,\Sigma)$, i.e, HETEROscedasticity

How was variance heterogeneity generated in these examples?

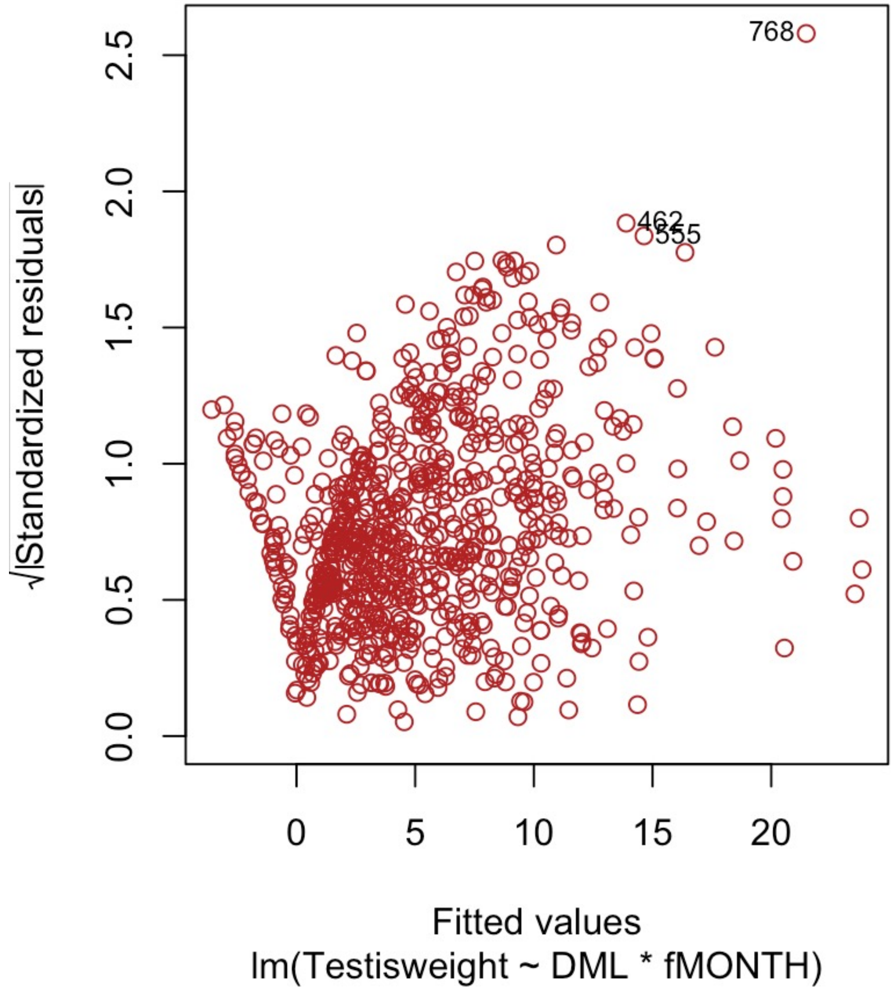
```

n=30
X = 1:n
e = rnorm(n,0,X)
Y = constant + slopeX * X + e
```



Goal: study seasonal patterns in reproductive and somatic tissues

Going back to the model of interest

$$\text{TestisWeight} = \text{constant} + \beta_1\text{DML} + \beta_2\text{Month} + \beta_3(\text{DML} \times \text{Mont}) + \ \text{e}$$



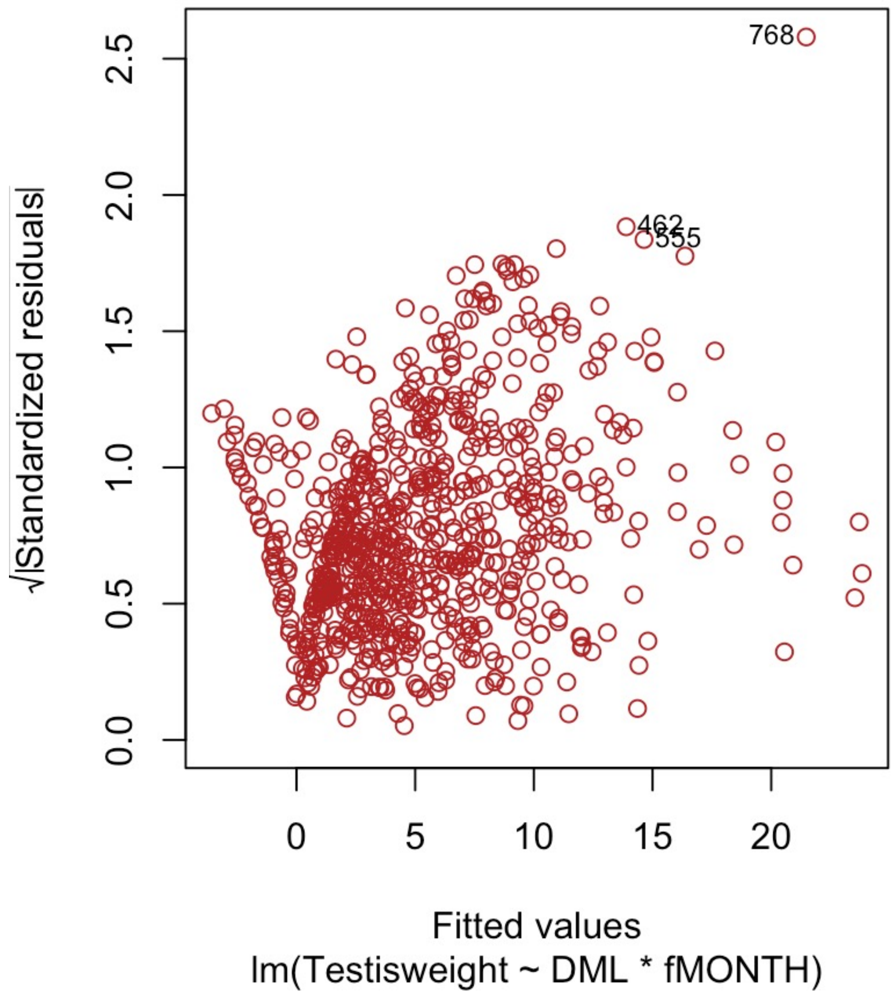
Residuals are highly heteroscedastic

```
> bptest(M1)

        studentized Breusch-Pagan test

data:  M1
BP = 160.08, df = 23, p-value < 2.2e-16
```

Goal: study seasonal patterns in reproductive and somatic tissues.

Going back to the model of interest

$$\text{TestisWeight} = \text{constant} + \beta_1\text{DML} + \beta_2\text{Month} + \beta_3(\text{DML} \times \text{Mont}) + \text{ e}$$



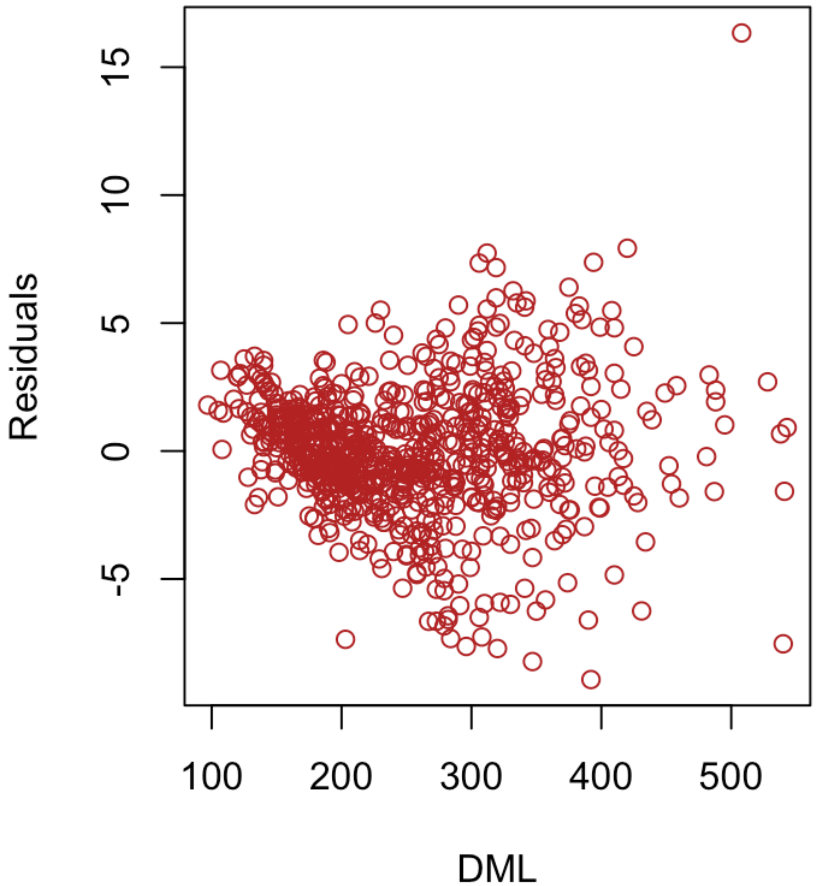
What are the origins (or proxies) of variation in residual variance?

```
> bptest(M1)

	studentized Breusch-Pagan test

data:  M1
BP = 160.08, df = 23, p-value < 2.2e-16
```

Goal: study seasonal patterns in reproductive and somatic tissues.

Variance changes as a function of DML

$$\mathrm{TestisWeight} = \mathrm{constant} + \beta_1 \mathrm{DML} + \beta_2 \mathrm{Month} + \beta_3 (\mathrm{DML} \times \mathrm{Month}) + \mathrm{e}$$

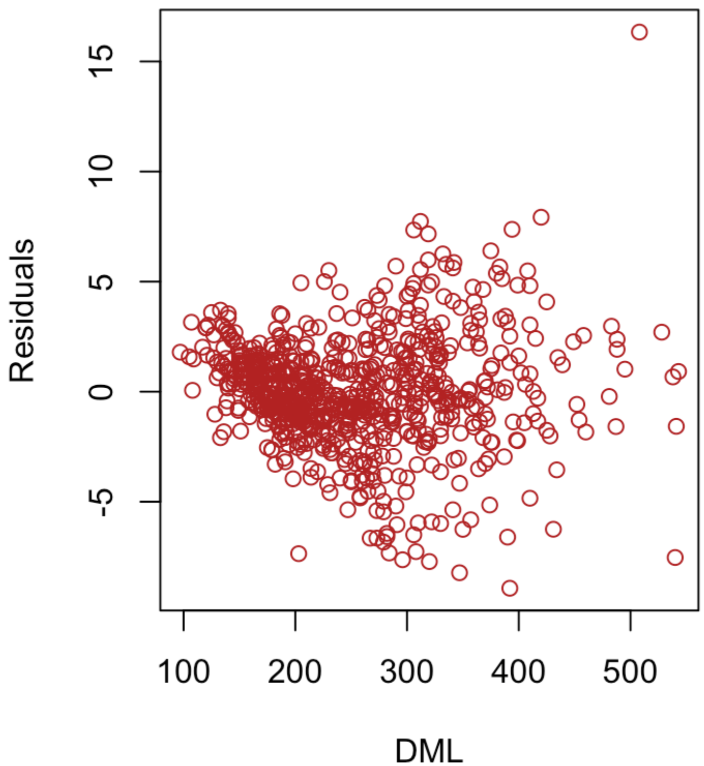


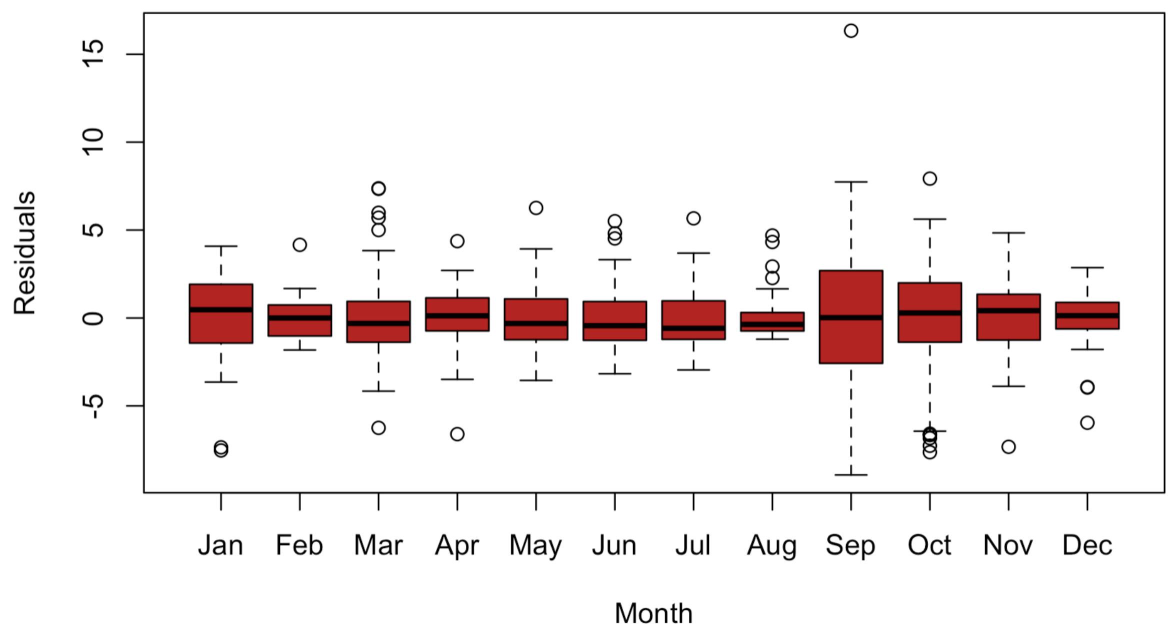
What are the origins (or proxies) of variation in residual variance?

Goal: study seasonal patterns in reproductive and somatic tissues.

Variance changes as a function of DML x Month (interaction)

$$\text{TestisWeight} = \text{constant} + \beta_1\text{DML} + \beta_2\text{Month} + \beta_3(\text{DML} \times \text{Month}) + \ \text{e}$$

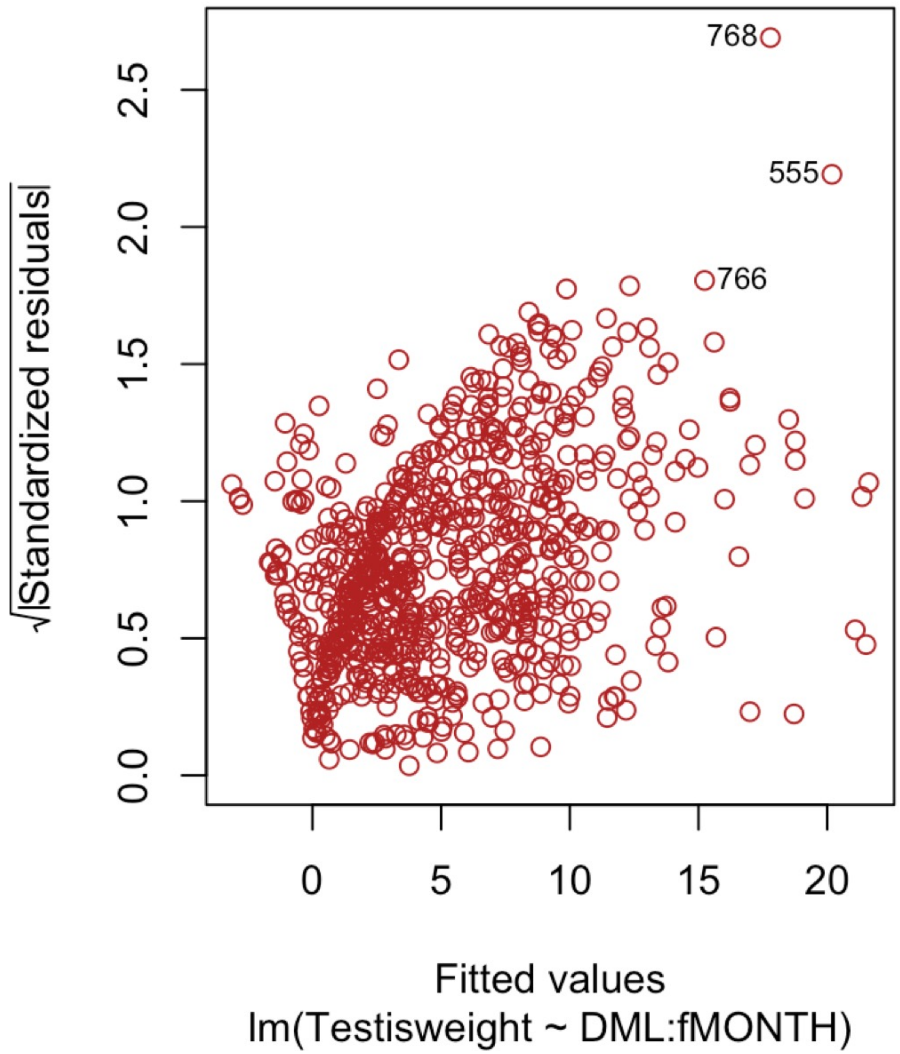




Goal: study seasonal patterns in reproductive and somatic tissues.

Variance changes as a function of DML x Month (interaction)

$$\text{TestisWeight} = \text{constant} + \beta_1 \text{DML} + \beta_2 \text{Month} + \beta_3 (\text{DML} \times \text{Month}) + \text{e}$$



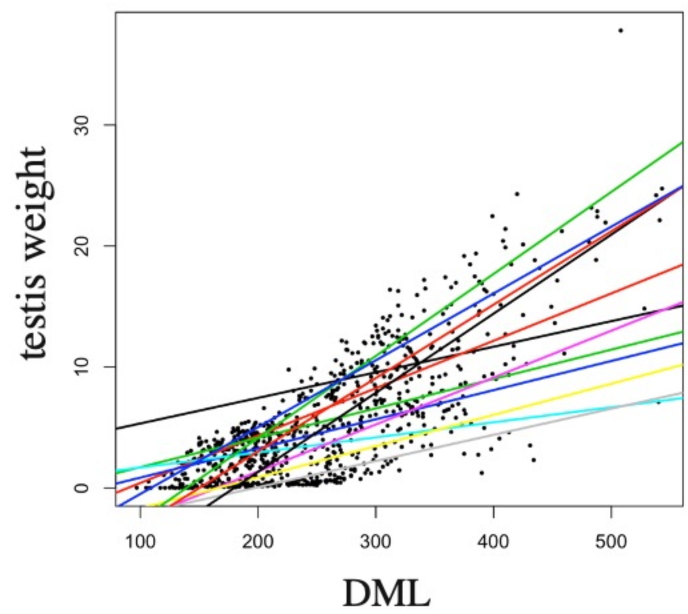
Variance changes as a function of Month

$$\text{TestisWeight} = \text{constant} + \beta_1\text{DML} + \beta_2\text{Month} + \beta_3(\text{DML} \times \text{Month}) + \text{e}$$

$$e \sim N(0, \sigma^2)$$ This assumption does not hold

If the DML by Month interaction is significant, we know that the slopes of DML change as a function of Month (i.e., ANCOVA).

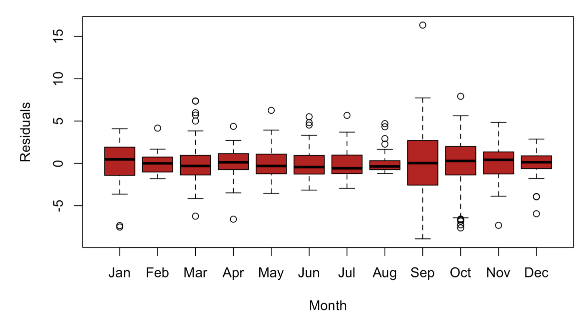
If the slopes for DML vary across months, assuming a single slope for all data will introduce heteroscedasticity. That is, residuals may be homoscedastic but only within models specific to each month.



Variance changes as a function of Month

$$e_{ij} \sim N(0, \sigma_j^2) \quad j = 1, \ldots, 12$$

$$\begin{array}{cc}
 & \begin{array}{cccc} \text{Specimen 1} & & \cdots & \text{Specimen 768} \end{array} \\
\begin{array}{c} \text{Specimen 1} \\ \vdots \\ \\ \text{Specimen 768} \end{array} & \underbrace{\begin{pmatrix} e_{ij} & 0 & \cdots & 0 \\ 0 & e_{ij} & \cdots & \vdots \\ \vdots & \cdots & e_{ij} & \vdots \\ 0 & \cdots & \cdots & e_{ij} \end{pmatrix}}_{\text{Variance-covariance matrix}}
\end{array}$$



Variance changes as a function of Month

$$e_{ij} \sim N(0, \sigma_j^2) \quad j = 1, \ldots, 12$$

How is this variance structure included in the model?

Ordinary Least Square GLS (fixed variance):

$$\beta = (X^{\mathrm{T}}X)^{-1}\, X^{\mathrm{T}}Y$$

Generalized Least Square GLS (variable variance):

$$\beta = (X^{\mathrm{T}}WX)^{-1}\, X^{\mathrm{T}}WY$$

How to account for variance differences?

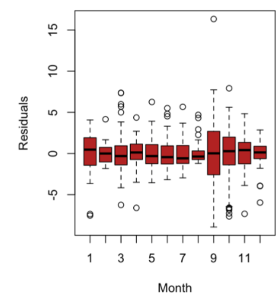
Variance changes as a function of Month

How is this variance structure included in the model?

Generalized Least Square GLS (variable variance):

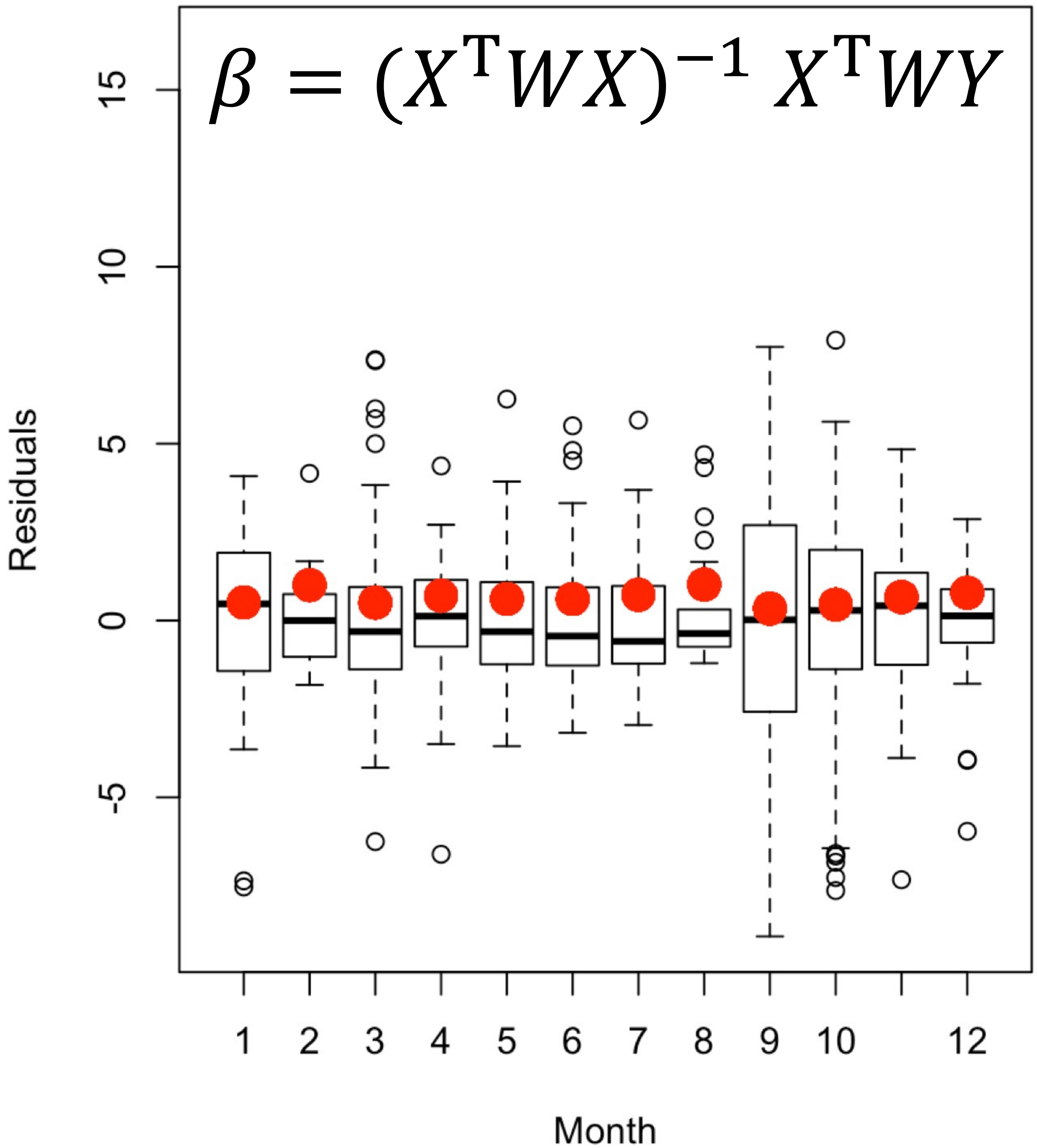
$$\beta = (X^{\mathrm{T}}WX)^{-1}\, X^{\mathrm{T}}WY \qquad W \sim 1/f(\Sigma)$$

$$\Sigma = \begin{array}{cc} & \begin{array}{cccc} \text{Specimen 1} & & \cdots & \text{Specimen 768} \end{array} \\ \begin{array}{c} \text{Specimen 1} \\ \vdots \\ \vdots \\ \text{Specimen 768} \end{array} & \underbrace{\begin{pmatrix} e_{ij} & 0 & \cdots & 0 \\ 0 & e_{ij} & \cdots & \vdots \\ \vdots & \cdots & e_{ij} & \vdots \\ 0 & \cdots & \cdots & e_{ij} \end{pmatrix}}_{\text{Variance-covariance matrix}} \end{array}$$

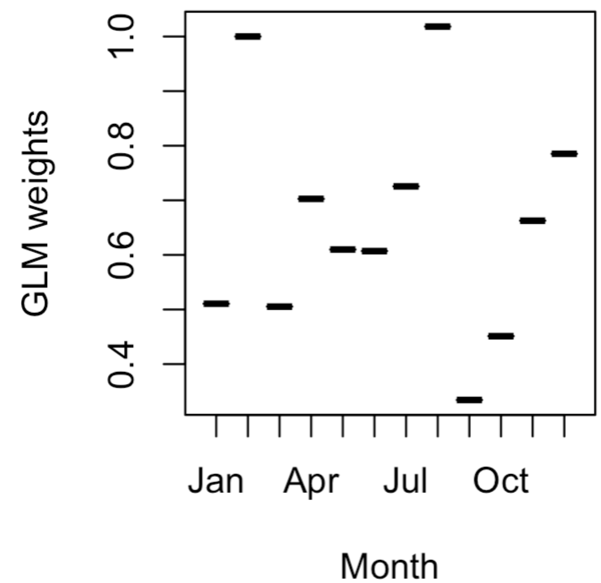


W is the reciprocal of a function of the variance-covariance matrix, but this function can take different forms (e.g., square root of residuals) or more complex structures. Using the reciprocal, specimens (within months here) with large residual will influence less the regression.

Variance changes as a function of Month & Weights are set inversely (reciprocal) to that variance



Weights

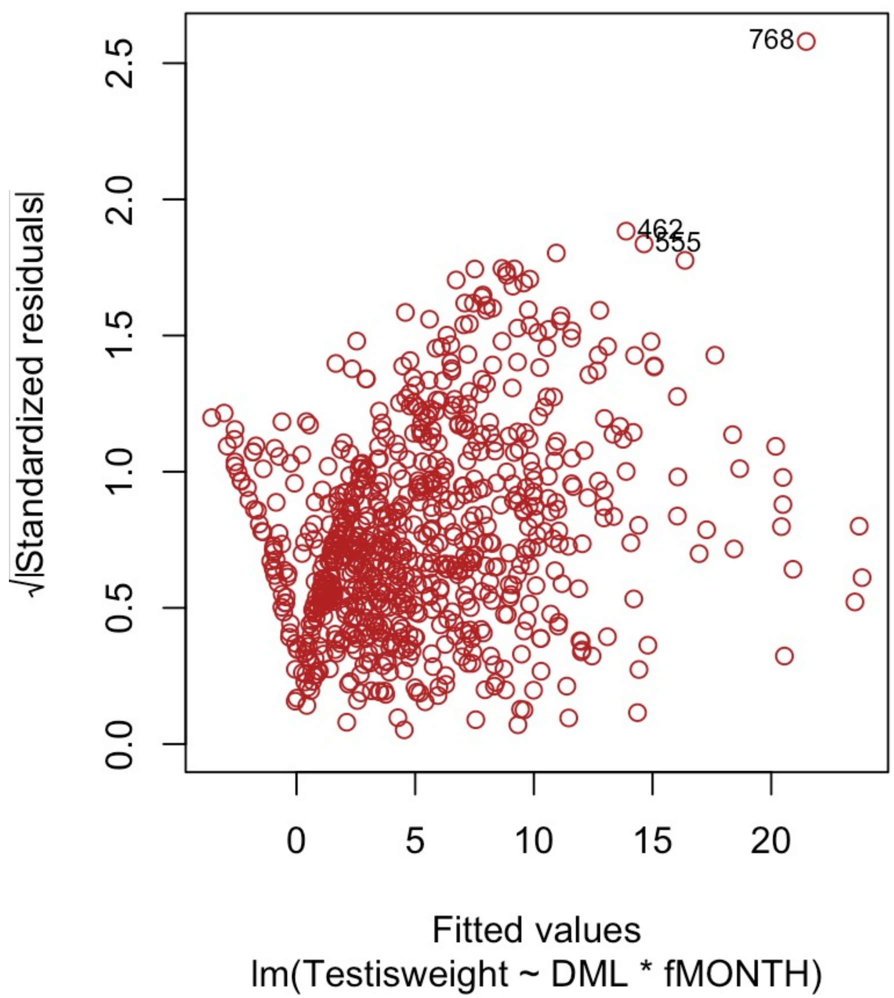


The weight of each individual is reciprocal to the residual variance of the month in which it was sampled.

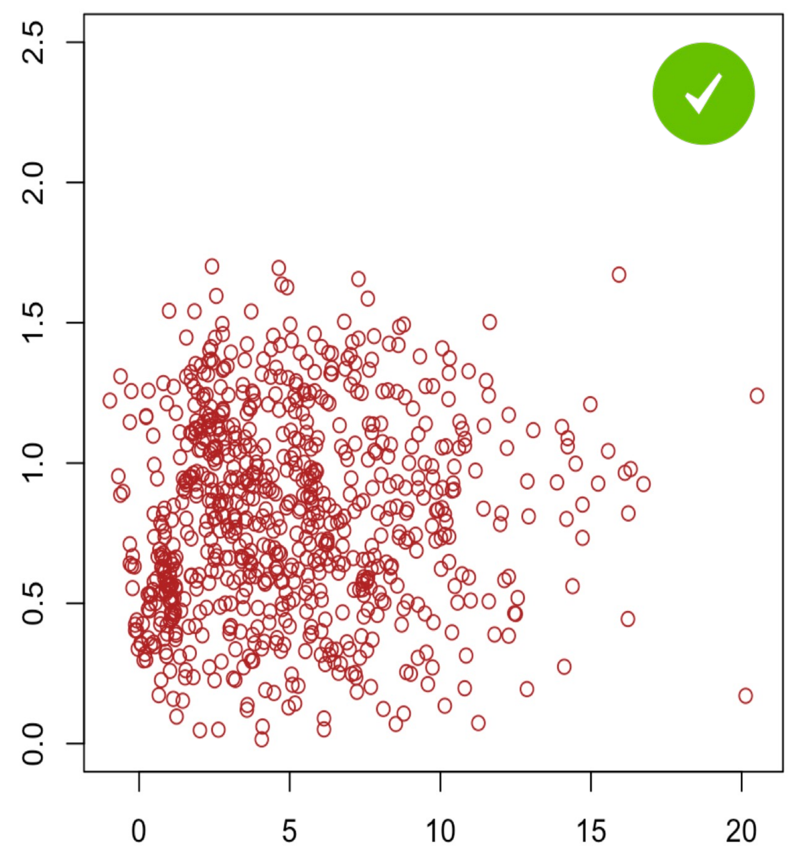
Goal: study seasonal patterns in reproductive and somatic tissues.

Contrasting OLS and GLS residual versus predicted plots

original model (OLS)



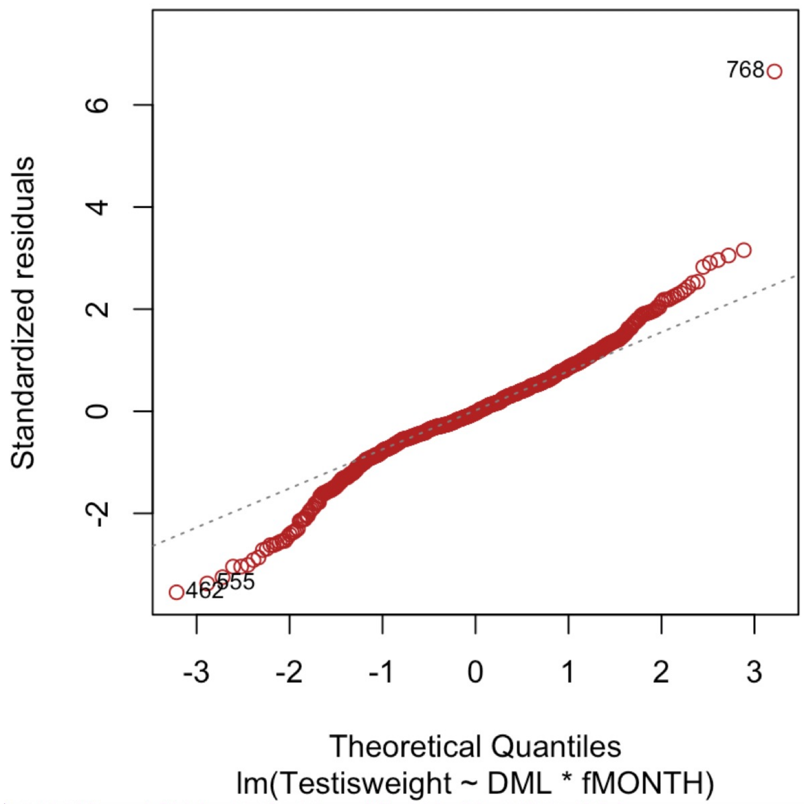
GLS Model



Goal: study seasonal patterns in reproductive and somatic tissues.

Q-Q normal residual plots

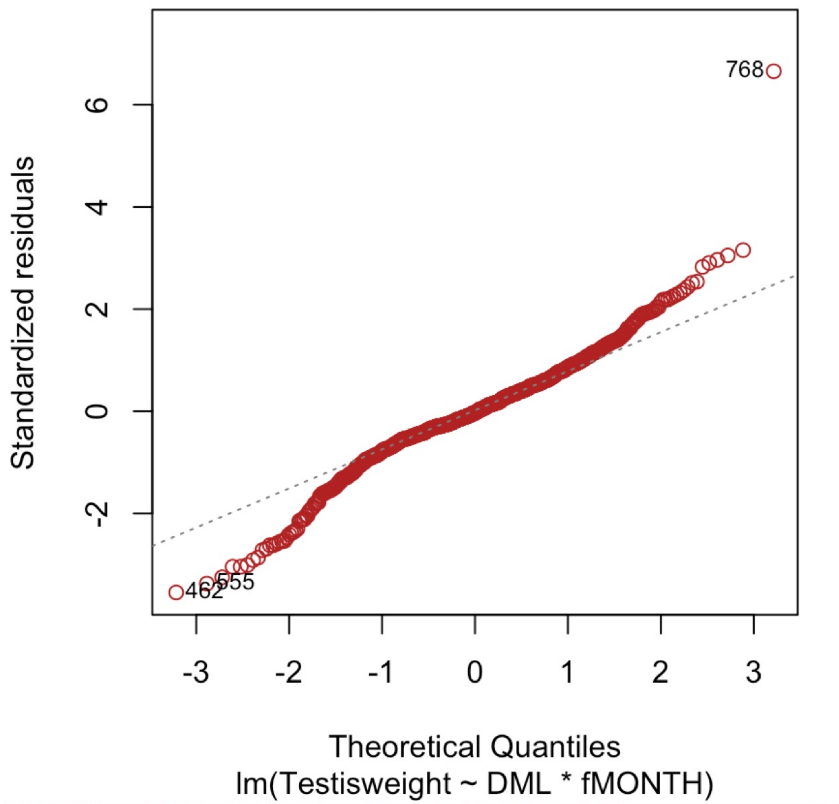
original model (OLS)



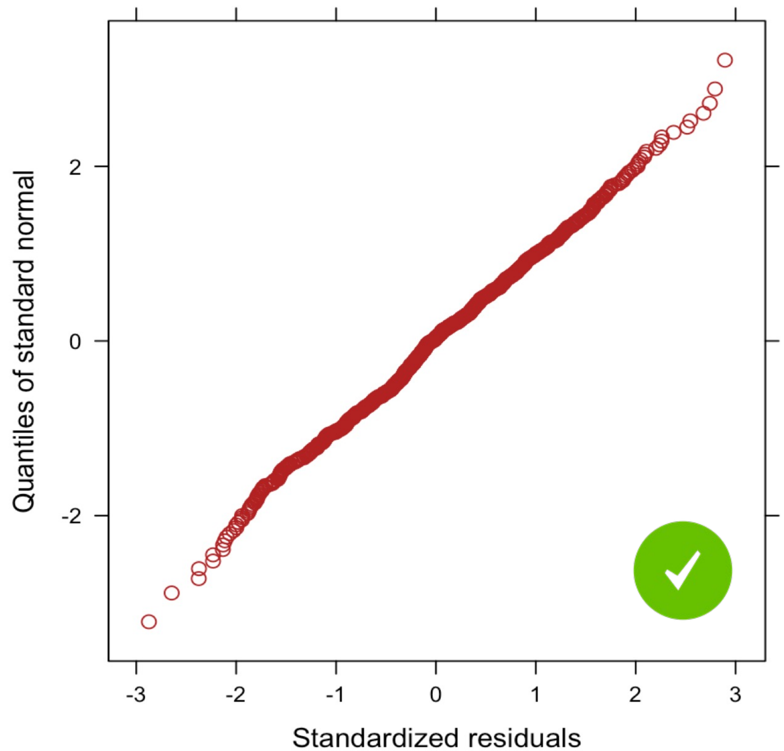
Goal: study seasonal patterns in reproductive and somatic tissues.

Q-Q normal residual plots

original model (OLS)



GLS Model



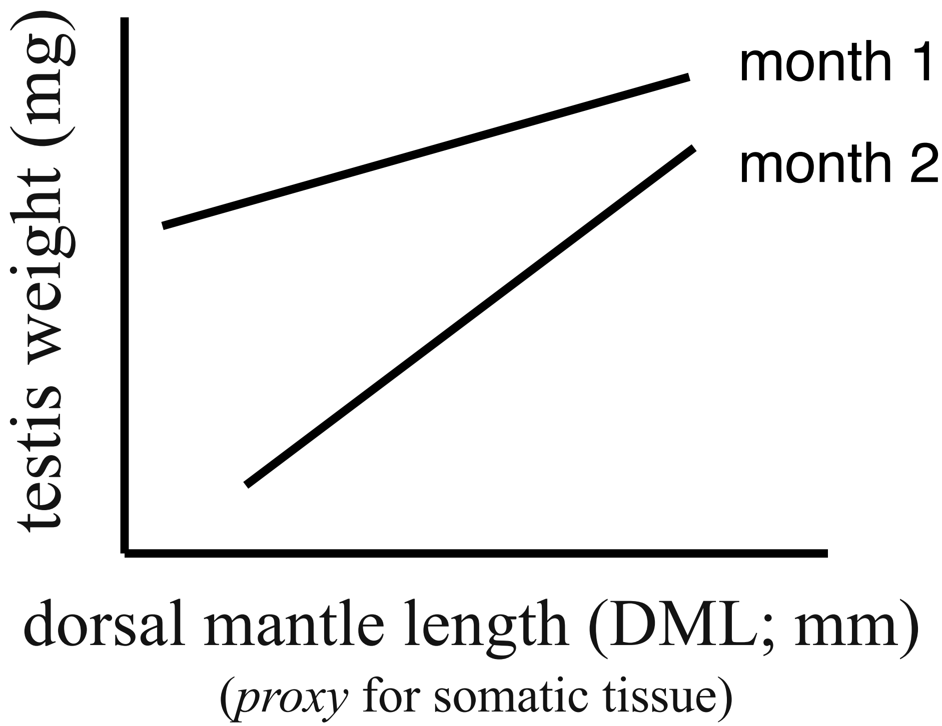
Seasonal patterns of investment in reproductive and somatic tissues in the squid *Loligo forbesi*

Jennifer M. Smith[1,a], Graham J. Pierce[1], Alain F. Zuur[2] and Peter R. Boyle[1]

[1] Department of Zoology, School of Biological Sciences, University of Aberdeen, Tillydrone Avenue, Aberdeen AB24 2TZ, UK
[2] Highland Statistics Ltd., 6 Laverock Road, Newburgh, Aberdeenshire, AB41 6FN, UK

Goal: study seasonal patterns in reproductive and somatic tissues.

In which month there is more investment (proportionally to amount of somatic tissues) in reproduction?



Aquat. Living Resour. 18, 341–351 (2005)

DOI: 10.1051/alr:2005038
www.edpsciences.org/alr

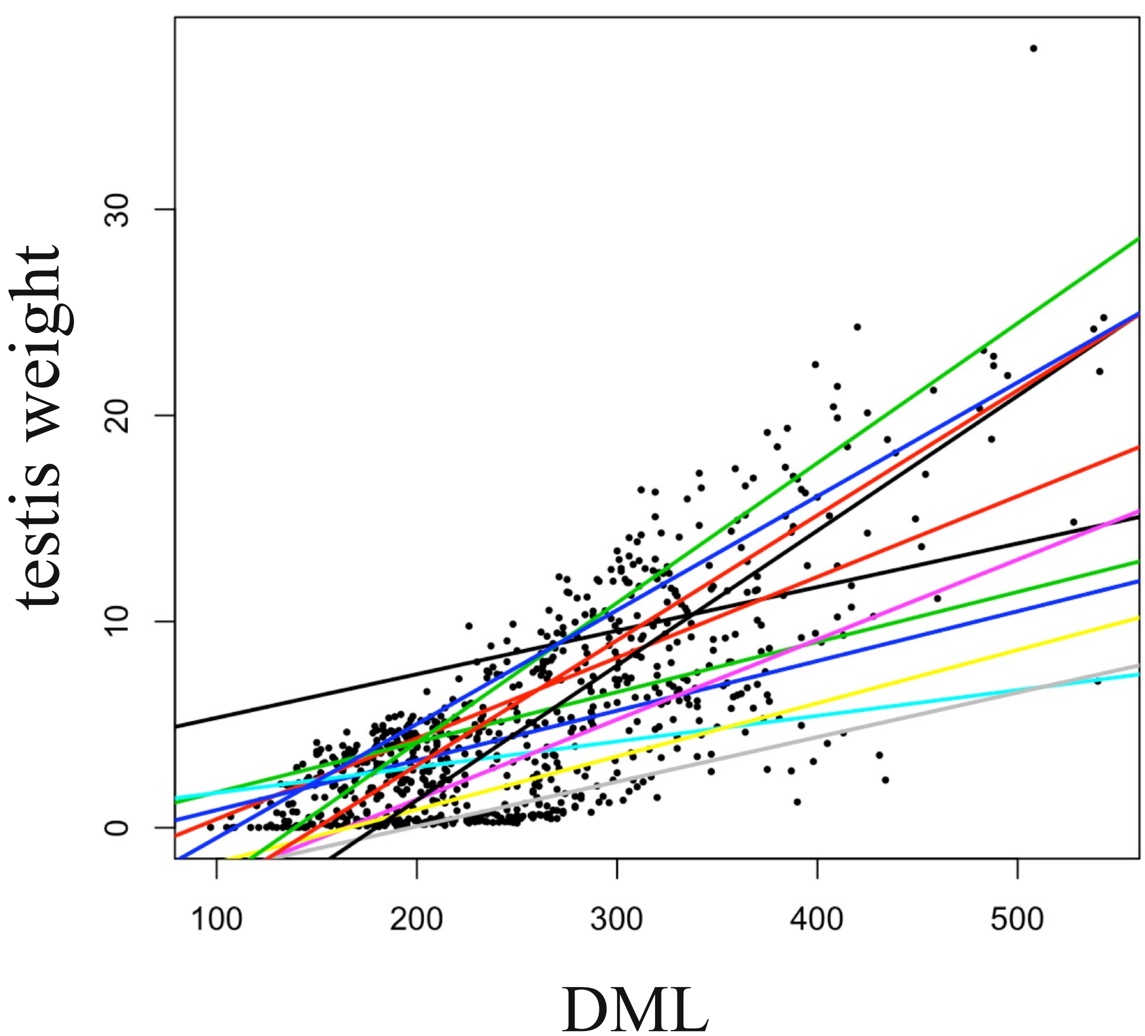
Goal: study seasonal patterns in reproductive and somatic tissues.

ANOVA results for GLS model

```
> anova(M.gls)
Denom. DF: 744
            numDF  F-value p-value
(Intercept)     1 3615.591  <.0001
DML             1 1648.534  <.0001
fMONTH         11   76.560  <.0001
DML:fMONTH     11   28.592  <.0001
```

$$\text{TestisWeight} = \text{constant} + \beta_1\text{DML} + \beta_2\text{Month} + \beta_3(\text{DML} \times \text{Month}) + \text{e}$$

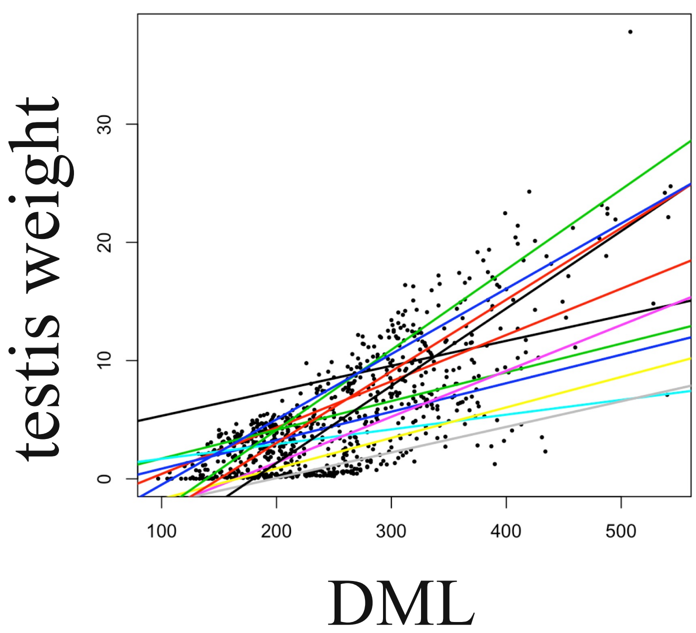
Interaction between dorsal mantle length (DML) and month indicating clear differences in reproductive investment among months (seasons)



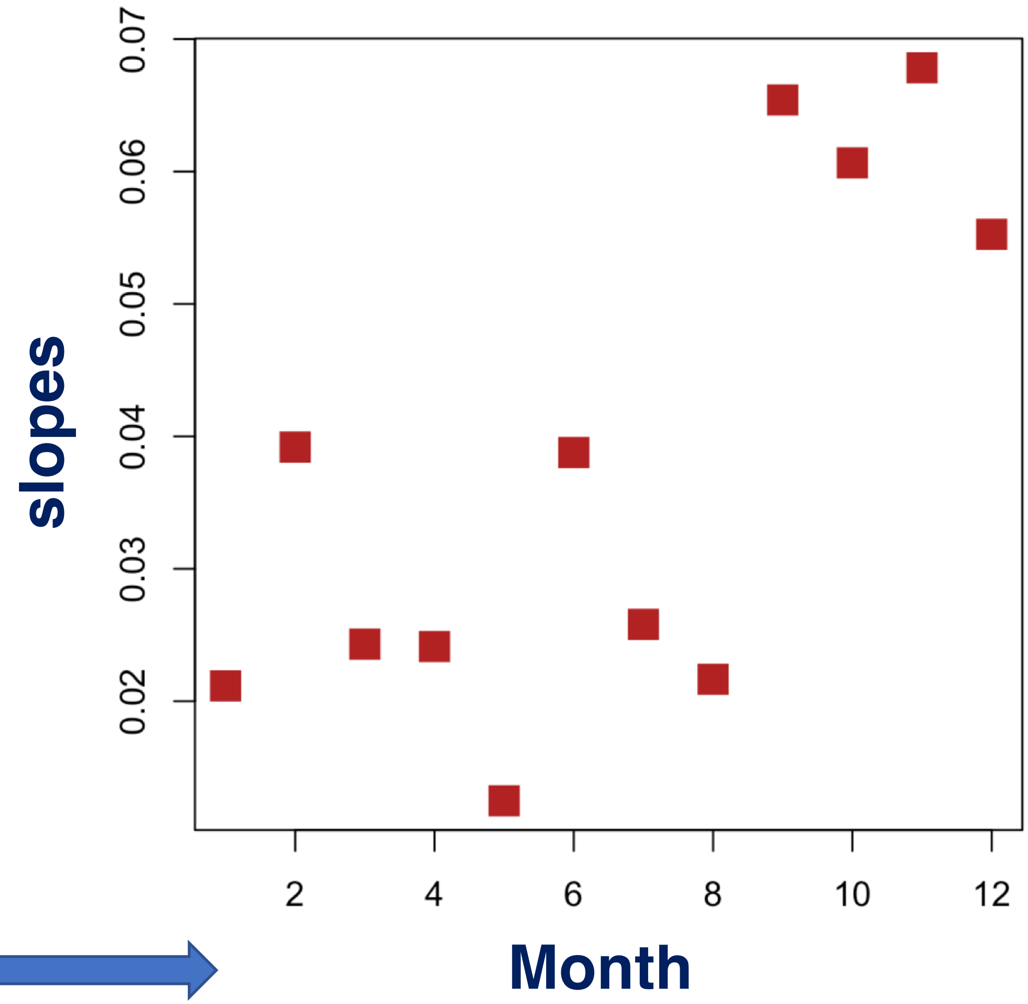
In which month there is more investment in reproduction?

```
> anova(M.gls)
Denom. DF: 744
            numDF  F-value p-value
(Intercept)     1 3615.591  <.0001
DML             1 1648.534  <.0001
fMONTH         11   76.560  <.0001
DML:fMONTH     11   28.592  <.0001
```

Interaction between dorsal mantle length (DML) and month indicating clear differences in reproductive investment among months (seasons)



How does investment change as a function of time?



Important points

There are many reasons and ways in which residual variance can change, as well as different types of functions (e.g., square root or more complex transformations) that can describe these changes.

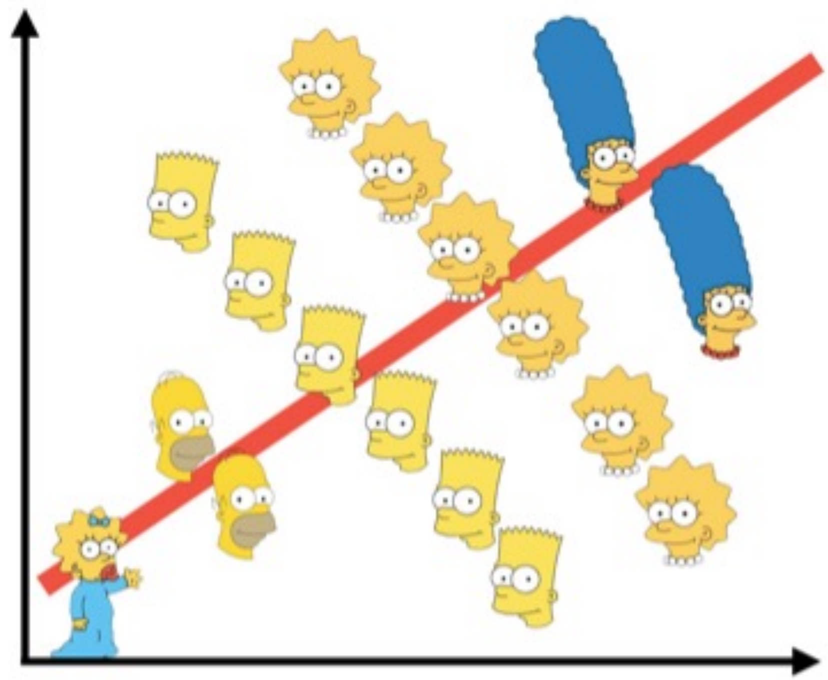
We can apply various structures and select the one that best fits the data (to be covered in the next lecture).

GLS, by itself, is not a mixed model—we will discuss this distinction in detail later. However, GLS is crucial for understanding variance heterogeneity and is often used within mixed-model frameworks.

The example explored here also allows understanding multiple slope variation (or parameter variation) which is essential to understand mixed models.

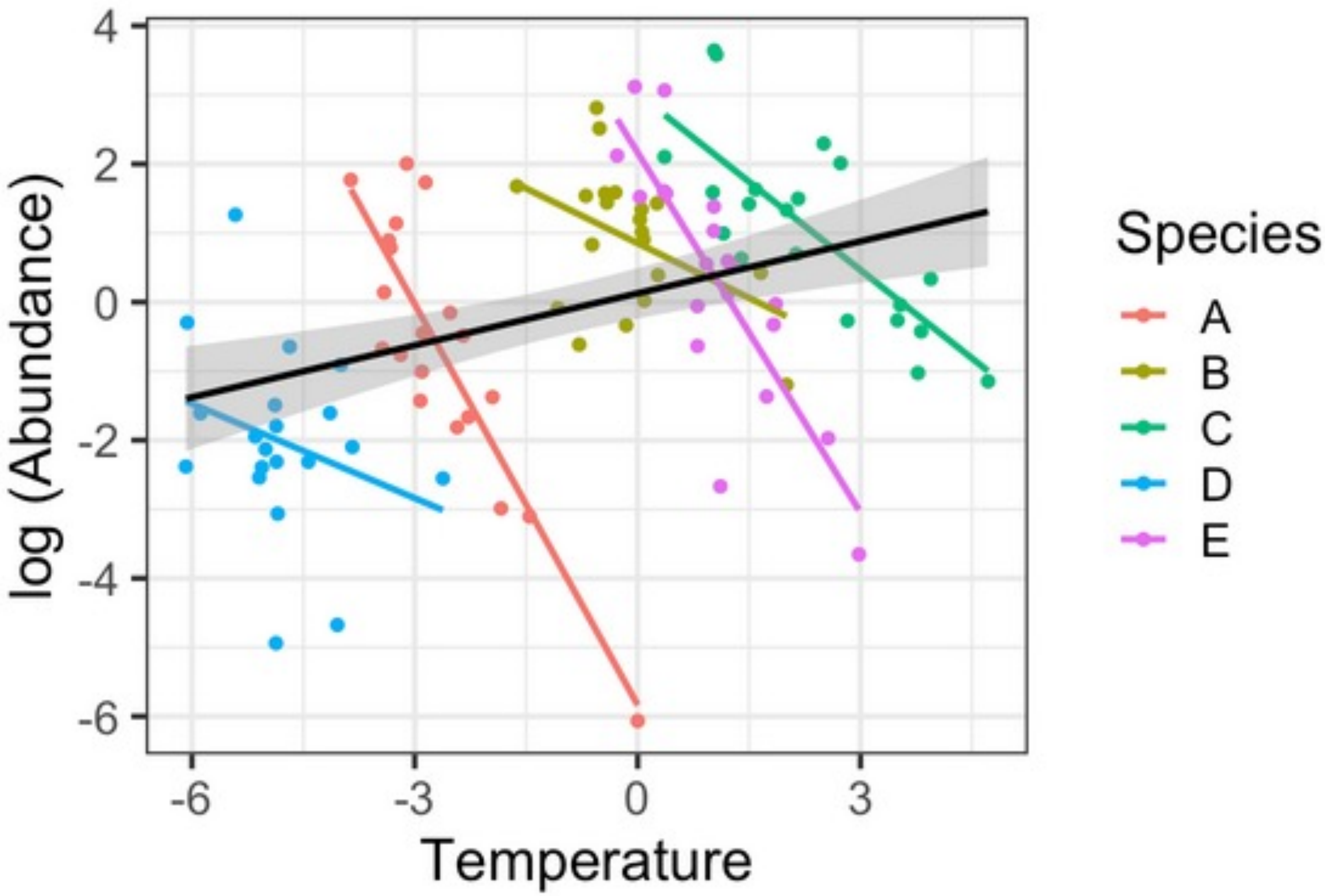
Next: a quick look into the general goals of a mixed model using Simpson’s paradox – more on mixed models in the next lectures.

“A phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined.”



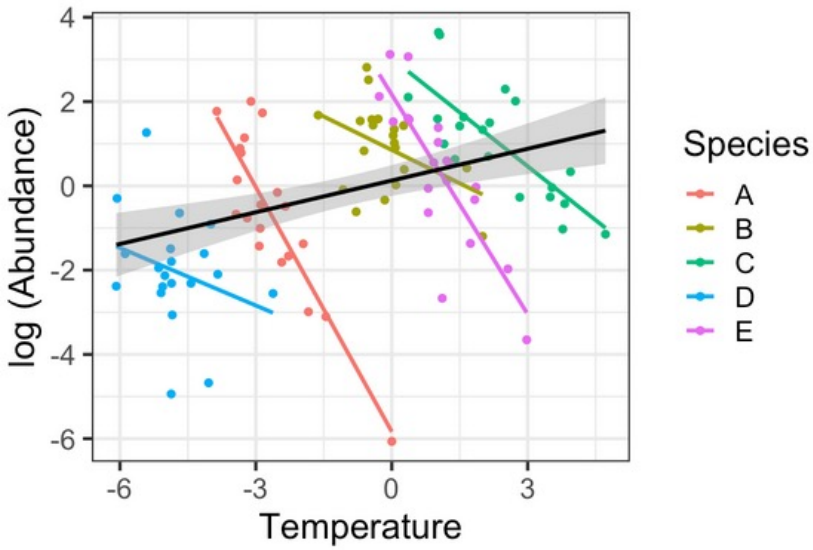
Important enough to have its own Wikipedia page: https://en.wikipedia.org/wiki/Simpson%27s_paradox

One feature in ecology is that species often differ in the way they respond to environmental variability. This can be well described by the Simpson's paradox (Simpson 1951), which is defined "as a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined."



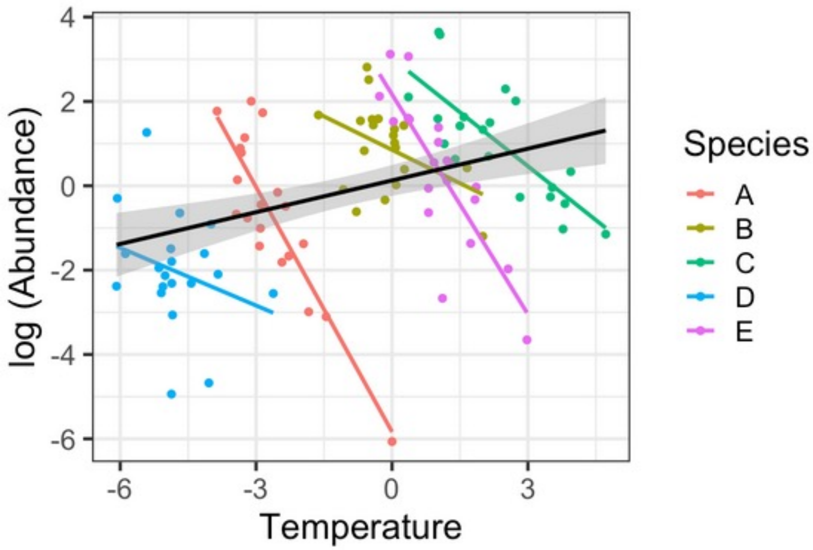
Important enough to have its own Wikipedia page:
https://en.wikipedia.org/wiki/Simpson%27s_paradox

Fixed effect model




```
## MODEL INFO:
## Observations: 100
## Dependent Variable: abundance
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,98) = 15.13, p = 0.00
## R² = 0.13
## Adj. R² = 0.12
##
## Standard errors: OLS
## ---------------------------------------------------------
##                              Est.   S.E.   t val.      p
## -------------------------- ------- ------ -------- ------
## (Intercept)                 -0.08   0.18    -0.48   0.63
## scale(temperature)           0.69   0.18     3.89   0.00
## ---------------------------------------------------------
##
## Continuous predictors are mean-centered and scaled by 1 s.d.
```

Mixed effect model




```
lm.mod.intercept <- lmer(abundance ~ temperature + (1|species),data=data.Simpson)
summ(lm.mod.intercept,scale = TRUE)
```

```
## MODEL INFO:
## Observations: 100
## Dependent Variable: abundance
## Type: Mixed effects linear regression
##
## MODEL FIT:
## AIC = 343.74, BIC = 354.16
## Pseudo-R² (fixed effects) = 0.30
## Pseudo-R² (total) = 0.95
##
## FIXED EFFECTS:
## ------------------------------------------------------------
##                        Est.    S.E.    t val.     d.f.      p
## ----------------- -------- ------- --------- -------- ------
## (Intercept)          -0.08    1.88     -0.04     3.77   0.97
## temperature          -2.84    0.34     -8.22    97.68   0.00
## ------------------------------------------------------------
##
```