Dealing with Heteroskedasticity

Michael Noonan

Biol 520C: Statistical modelling for biological data

Table of contents

Heteroskedasticity and the IID Assumption

The models so far

We started with simple linear regression:

$$y_i = \beta_0 + \beta x_i + \varepsilon_i$$

We then extended this to multiple linear regression of the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \varepsilon_i$$

And then to linear mixed effects model that account for correlations within groups:

$$\boldsymbol{y_i} = X_i \boldsymbol{\beta} + Z_i \boldsymbol{b_i} + \boldsymbol{\varepsilon_i}$$

But the elephant in the room is our assumption that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

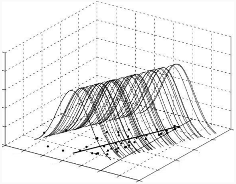
The IID Assumption

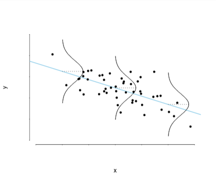
The models we've been working with so far assume that the residuals at each x_i value should be Independent and Identically Distributed (IID).

More specifically, they should be normally distributed with a constant variance σ^2:

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

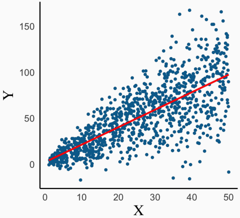
In real systems variances can differ across groups, times, etc. (i.e., heteroskedasticity) and the 'identical' part of the IID assumption is likely to be broken in many cases.





Heteroskedasticity

Wikipedia: In statistics, a vector of random variables is heteroskedastic (from Ancient Greek hetero ‘different’ and skedasis ‘dispersion’) if the variability of the random disturbance is different across elements of the vector.



Formally, for a hypothetical dataset with three observations , A is homoskedastic ; in B, the variance increases steadily across samples ; in C, the variance depends on the value of x.

$$A = \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad B = \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \qquad C = \sigma^2 \begin{bmatrix} x_1 & 0 & 0 \\ 0 & x_2 & 0 \\ 0 & 0 & x_3 \end{bmatrix}$$

More generally, if the variance-covariance matrix of ε_i across i has a non-constant diagonal, the errors will be heteroskedastic.

Heteroskedasticity cont.

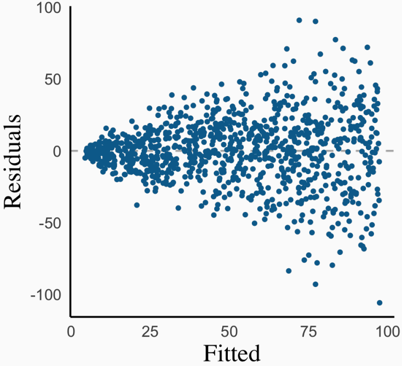
What heteroskedasticity boils down to is non-constant variance that varies with other components of the system.

When the IID assumption is violated by heteroskedasticity variances, and hence standard errors, will be poorly estimated resulting in misleading *p*-values (we'll explore this concept in this week's practical).

The deterministic part of the model *may* be correctly specified, but any predictions are likely to be off.

Identifying heteroskedasticity

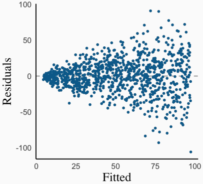
The easiest way to identify the presence of heteroskedasticity is by plotting a model's residuals against the predicted values.



These cone shaped residuals are the tell-tale sign of heteroskedasticity.

Identifying the cause of heterosc.

Identify the presence of heteroskedasticity is fairly easy.



In order to deal with the issue we need to identify the cause (i.e., what group/factor does the variance change with).

The easiest way to identify the cause of heteroskedasticity is by plotting a model's residuals against each of the predictors.

The squid data

Today we're going to work with a dataset collected by Smith *et al.* (2005) to examine seasonal patterns in reproductive and somatic tissues in the squid *Loligo forbesi*.

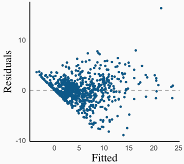
The data are comprised of measures of testis weight (TW) and dorsal mantle length (DML) from 768 male squid. The month of data collection was also recorded.

Today's starting point is the linear regression model:

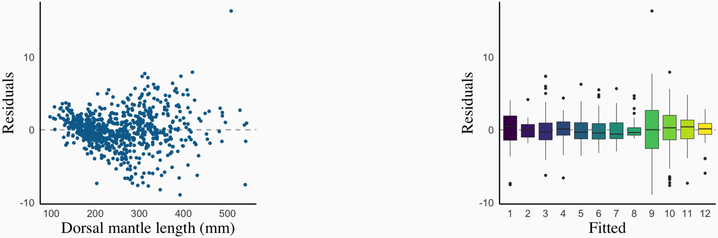
$$\mathrm{TW}_i = \beta_0 + \beta_1 \mathrm{DML}_i + \beta_2 \mathrm{month}_i + \beta_3 \mathrm{DML}_i : \mathrm{month}_i + \varepsilon_i$$

Heteroskedasticity in the squid data

```
library(nlme)

data <- read.csv("Squid.csv")

data$month <- as.factor(data$month)

mod <- gls(TW ~ DML*month, data = data)
```



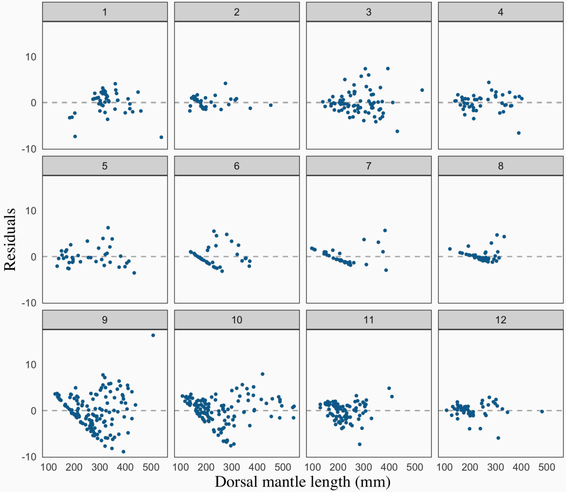
These data are clearly heteroskedastic. To identify the cause we'll look at the model's residuals against DML and month.



$>$DML $= >\sigma^2$, but different months also have different σ^2.

The source

Residual spread changes month to month, and in some months increases with DML. So both need to be modelled for a full correction.



Heterosc. in the squid data cont

```
summary(mod)

                 Value Std.Error   t-value p-value
(Intercept)   3.215222 1.8740686  1.715637  0.0866
DML           0.021157 0.0055585  3.806244  0.0002
month2       -6.720020 2.3282175 -2.886337  0.0040
month3       -3.926923 2.2081953 -1.778340  0.0758
month4       -4.772045 2.2875351 -2.086108  0.0373
month5       -2.771457 2.2656715 -1.223239  0.2216
month6       -9.598061 2.4163109 -3.972196  0.0001
month7       -7.494959 2.2933137 -3.268178  0.0011
month8       -7.479426 2.9310875 -2.551758  0.0109
month9      -14.963009 2.0353133 -7.351698  0.0000
month10     -12.320827 1.9637718 -6.274063  0.0000
month11     -12.650490 2.1922272 -5.770611  0.0000
month12      -9.235813 2.2264505 -4.148223  0.0000
DML:month2    0.018032 0.0083101  2.169894  0.0303
DML:month3    0.003151 0.0068657  0.458989  0.6464
DML:month4    0.002972 0.0074316  0.399848  0.6894
DML:month5   -0.008677 0.0072378 -1.198801  0.2310
DML:month6    0.017620 0.0084180  2.093183  0.0367
DML:month7    0.004647 0.0079929  0.581432  0.5611
DML:month8    0.000500 0.0105569  0.047373  0.9622
DML:month9    0.044242 0.0062130  7.120835  0.0000
DML:month10   0.039495 0.0059673  6.618449  0.0000
DML:month11   0.046671 0.0074640  6.252829  0.0000
DML:month12   0.034099 0.0077022  4.427127  0.0000
```

Running summary on the fit shows a lot of significant parameters.

But because of the heteroskedasticity we can't trust the *p*-values.

We could try to transform the data, but the heteroskedasticity is interesting ecological and we should avoid treating it as a 'nuisance' and try to model it.

Variance Structures

Fixed Variance

One of the simplest variance structures is the so called ‘fixed variance structure’.

$$\varepsilon_i \sim \mathcal{N}(0,\, \sigma^2 \times x_i)$$

This formulation allows for variance to increase with x

Because σ^2 is already being estimated, and x_i are our data, there are no extra parameters to fit with this formulation.

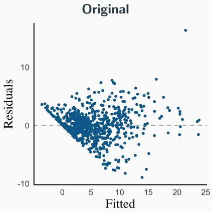
In other words, we pick this structure up for free, which is big bonus, but it also means we don’t learn anything new about the system.

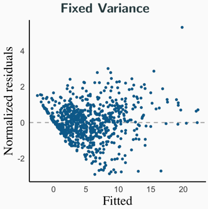
Fixed Variance in R

For our squid data, variance increased with DML, so a reasonable formulation to handle the heteroskedasticity would be: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \times \mathrm{DML}_i)$

We can do this via the `weights` argument in the `gls()` function.

```
mod2 <- gls(TW ~ DML*month,
            weights = varFixed(~DML),
            data = data)

anova(mod, mod2)

     Model df      AIC      BIC    logLik
mod      1 25 3752.084 3867.385 -1851.042
mod2     2 25 3620.898 3736.199 -1785.449
```





AIC shows an improvement, but the residuals still show signs of heterosk.

Constant Variance

Another way for heteroskedasticity to enter into the system is for different **categorical** groups to have different variances (e.g., males/females, seasons, years, etc...).

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$$

If there is only one group, the variance is constant. When there is a grouping factor with $j > 1$ levels, the variance function allows j different variances, one for each level of the factor.

In practice, the variance function represents the ratios between the different σ^2 and a reference σ^2, so $j - 1$ new coefficients are needed to represent the variance function.

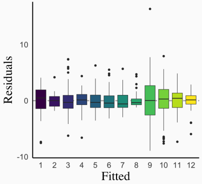
$$\sigma_1^2 = \sigma^2 \times 1, \quad \sigma_2^2 = \sigma^2 \times \theta_1 \quad \dots \quad \sigma_j^2 = \sigma^2 \times \theta_{j-1}$$

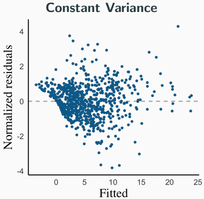
Constant Variance in R

For our squid data, variance differed across months, so we can have the variance change across months: $\varepsilon_{i,month} \sim \mathcal{N}(0, \sigma^2_{month})$

We can do this via the `varIdent()` function.

```
mod3 <- gls(TW ~ DML*month,
            weights = varIdent(form = ~1|month),
            data = data)

summary(mod3)
...
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | month
 Parameter estimates:
        2         9        12        11         8
1.0000000 2.9913585 1.2736165 1.5090469 0.9821378
       10         5         7         6         4
2.2162169 1.6396315 1.3783514 1.6473098 1.4232364
        1         3
1.9584902 1.9788664
...
anova(mod, mod3)

     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod      1 25 3752.084 3867.385 -1851.042
mod3     2 36 3614.436 3780.469 -1771.218 1 vs 2 159.6479  <.0001
```





Again, AIC shows an improvement, but the residuals aren't convincing.

Power Variance

The third variance structure we'll look at is the power variance structure:

$$\varepsilon_{ij} \sim \mathcal{N}(0,\, \sigma^2 \times |x_i|^{2\delta_j})$$

The power variance structure is a generalisation of the fixed variance structure.

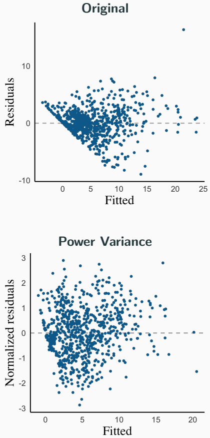
- If $\delta = 0$ this reduces to the linear regression model.
- If $\delta = \frac{1}{2}$ this reduces to the fixed variance structure.
- If $\delta > \frac{1}{2}$ the variance is allowed to spread more than a fixed variance for the same increase in x.
- When a grouping factor is present, a different δ is used for each factor level.

Power Variance in R

For our squid data, variance increased with DML, so a reasonable formulation would be: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \times |DML_i|^{2\delta})$

We can do this via the `varPower()` function.

```
mod4 <- gls(TW ~ DML*month,
            weights = varPower(form = ~ DML),
            data = data)
 summary(mod4)
...
Variance function:
 Structure: Power of variance covariate
 Formula: ~DML
 Parameter estimates:
   power
1.759009
...
anova(mod, mod4)

     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod      1 25 3752.084 3867.385 -1851.042
mod4     2 26 3473.019 3592.932 -1710.509 1 vs 2 281.0648  <.0001
```

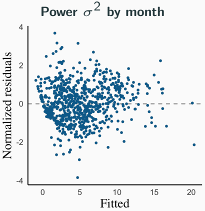


AIC shows a big improvement, and the residuals look much better.

Power Variance in R cont.

Variance increased with DML and across months, so we could also try $\varepsilon_{i,\text{month}} \sim \mathcal{N}(0,\, \sigma^2 \times |DML_i|^{2\delta_{\text{month}}})$

```
mod5 <- gls(TW ~ DML*month,
            weights = varPower(form = ~ DML|month),
            data = data)

summary(mod5)

...
Variance function:
 Structure: Power of variance covariate, different strata
 Formula: ~DML | month
 Parameter estimates:
       2        9       12       11        8       10
1.728531 1.789503 1.733561 1.749264 1.617694 1.789187
       5        7        6        4        1        3
1.746517 1.673244 1.754488 1.711371 1.698945 1.722742
...
anova(mod, mod5)

     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod      1 25 3752.084 3867.385 -1851.042
mod5     2 37 3407.511 3578.156 -1666.755 1 vs 2 368.5728  <.0001
```



Best AIC yet, and the residuals look pretty good.

Exponential Variance

The next variance structure we'll look at is the exponential variance structure:

$$\varepsilon_{ij} \sim \mathcal{N}(0,\, \sigma^2 \times e^{2\delta_j \times x_i})$$

The exponential variance structure models the variance of the residuals as σ^2 multiplied by an exponential function.

- If $\delta = 0$ this reduces to the linear regression model.
- If $\delta > 0$ the variance increases with x.
- If $\delta < 0$ the variance decreases with x.
- As with the power variance, δ can be grouped by a categorical variable j.

Exponential Variance in R

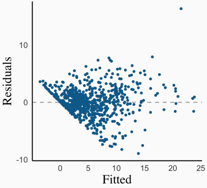
For our squid data, variance increased with DML, so our formulation is:

$$\varepsilon_i \sim \mathcal{N}(0,\, \sigma^2 \times e^{2\delta \times DML_i})$$

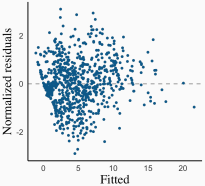
We can implement this via the `varExp()` function.

```
mod6 <- gls(TW ~ DML*month,
              weights = varExp(form = ~ DML),
              data = data)
 summary(mod6)
...
Variance function:
 Structure: Exponential of variance covariate
 Formula: ~DML
 Parameter estimates:
      expon
0.006832722
...
anova(mod, mod6)

     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod      1 25 3752.084 3867.385 -1851.042
mod6     2 26 3478.152 3598.066 -1713.076 1 vs 2 275.9312  <.0001
```

Original



Exponential Variance

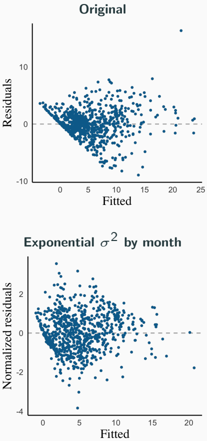


AIC shows an improvement, and the residuals look ok.

Exponential Variance in R cont

Again, we can assign different variances for different months via the `varExp()` function.

```
mod7 <- gls(TW ~ DML*month,
            weights = varExp(form = ~ DML|month),
            data = data)
 summary(mod7)
...
Variance function:
 Structure: Exponential of variance covariate, different strata
 Formula: ~DML | month
 Parameter estimates:
          2           9          12          11           8
0.005463411 0.007471853 0.006089471 0.006700889 0.004340133
         10           5           7           6           4
0.007371252 0.006060562 0.005575366 0.006959413 0.005612065
          1           3
0.005510245 0.006214627
...
anova(mod, mod7)

     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod      1 25 3752.084 3867.385 -1851.042
mod7     2 37 3419.719 3590.364 -1672.859 1 vs 2 356.3652  <.0001
```



AIC shows an improvement, and the residuals look good.

Constant Plus Power Variance

The penultimate variance structure we'll look at is the constant plus power variance structure:

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \times (\delta_{1j} + |x_i|^{\delta_{2j}})^2)$$

- If $\delta_1 = 0$ & $\delta_2 = 0$ this reduces to the linear regression model.
- If $\delta_1 = 0$ & $\delta_2 \neq 0$ this reduces to the power variance.
- Here δ_1 & δ_2 can be grouped by a categorical variable j.

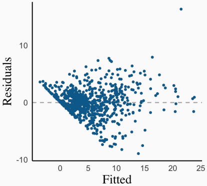
Constant Plus Power Variance in R

For our squid data our formulation is: $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \times (\delta_1 + |DML_i|^{\delta_2})^2)$

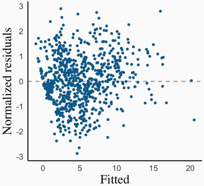
Which we implement via `varConstPower()`.

```
mod8 <- gls(TW ~ DML*month,
            weights = varConstPower(form = ~ DML),
            data = data)
 summary(mod8)
...
Variance function:
 Structure: Constant plus power of variance covariate
 Formula: ~DML
 Parameter estimates:
    const     power
0.0992986 1.7590192
...
anova(mod, mod8)

     Model df      AIC      BIC    logLik   Test L.Ratio p-value
mod      1 25 3752.084 3867.385 -1851.042
mod8     2 27 3475.019 3599.544 -1710.509 1 vs 2 281.065  <.0001
```

Original



Constant Plus Power Variance

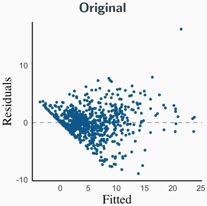


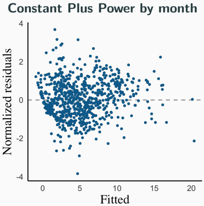
AIC shows an improvement, and the residuals look ok.

Exponential Variance in R cont

Again, we can assign different variances for different months via the `varConstPower()` function.

```
mod9 <- gls(TW ~ DML*month,
            weights = varConstPower(form = ~ DML|month),
            data = data)
 summary(mod9)
...
Variance function:
 Structure: Constant plus power of variance covariate, different
       strata
 Formula: ~DML | month
 Parameter estimates:
              2         9        12          11           8
const 0.1098482 0.1044132 0.0978525 0.09883092 0.09952452
power 1.7285761 1.7895481 1.7336082 1.74931089 1.61774116
              10          5           7           6          4
const 0.09361078 0.09794217 0.09729915 0.09819052 0.0975102
power 1.78923387 1.74656454 1.67329035 1.75453620 1.7114162
              1         3
const 0.0988839 0.09964889
power 1.6989915 1.72278732
...
anova(mod, mod9)

     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod      1 25 3752.084 3867.385 -1851.042
mod9     2 49 3431.511 3657.501 -1666.755 1 vs 2 368.5728  <.0001
```





AIC and residuals look good, but note all the DFs!

Combination of Variances

The last variance structure we'll look at is the combined variance structure.

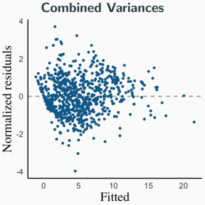
This is an `R` function (`varComb`) as opposed to a specific formulation, and allows you to flexibly combine the previously described structures.

The corresponding variance function is equal to the product of the variance functions.

It's very flexible, but challenging to work with.

Combination of Variances in R

```
mod10 <- gls(TW ~ DML*month,
             weights = varComb(varIdent(form = ~1|month),
                                 varExp(form = ~DML)),
             data = data)
summary(mod10)

...
Combination of variance functions:
 Structure: Different standard deviations per stratum
 Formula: ~1 | month
 Parameter estimates:
        2         9        12        11         8        10
1.0000000 1.4806117 1.0577473 1.1948695 0.5829504 1.4255755
        5         7         6         4         1         3
1.1276263 0.7353367 1.2712401 0.9710625 0.8668377 1.0561607
 Structure: Exponential of variance covariate
 Formula: ~DML
 Parameter estimates:
      expon
0.006711743
...
anova(mod, mod10)

      Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod       1 25 3752.084 3867.385 -1851.042
mod10     2 37 3414.817 3585.463 -1670.409 1 vs 2 361.2663  <.0001
```



Selecting the best Structure

We just fit 10 different models, but how do we know which variance structure to go with?

Not all variance structures are nested, so the easiest way to pick the best one is by AIC based model selection.

```
#Calculate AIC values
TABLE <- AIC(mod, mod2, mod3, mod4,
             mod5, mod6, mod7, mod8,
             mod9, mod10)

#Ordered by lowest to highest AIC
TABLE <- TABLE[order(TABLE$AIC),]

#Calculate Delta AICs
TABLE$DeltaAIC <- TABLE$AIC - TABLE$AIC[1]

#Evidence compared to AIC best model
TABLE$Evidence <- 1/exp(-TABLE$DeltaAIC/2)
```

```
      df      AIC   DeltaAIC     Evidence
mod5  37 3407.511   0.000000 1.000000e+00
mod10 37 3414.817   7.306579 3.860144e+01
mod7  37 3419.719  12.207673 4.475717e+02
mod9  49 3431.511  24.000039 1.627580e+05
mod4  26 3473.019  65.507991 1.678338e+14
mod8  27 3475.019  67.507874 4.561929e+14
mod6  26 3478.152  70.641592 2.185883e+15
mod3  36 3614.436 206.924992 8.574149e+44
mod2  25 3620.898 213.386607 2.169267e+46
mod   25 3752.084 344.572838 6.653361e+74
```

Corrected model

Original Model

```
                  Value Std.Error   t-value p-value
(Intercept)    3.215222 1.8740686  1.715637  0.0866
DML            0.021157 0.0055585  3.806244  0.0002
month2        -6.720020 2.3282175 -2.886337  0.0040
month3        -3.926923 2.2081953 -1.778340  0.0758
month4        -4.772045 2.2875351 -2.086108  0.0373
month5        -2.771457 2.2656715 -1.223239  0.2216
month6        -9.598061 2.4163109 -3.972196  0.0001
month7        -7.494959 2.2933137 -3.268178  0.0011
month8        -7.479426 2.9310875 -2.551758  0.0109
month9       -14.963009 2.0353133 -7.351698  0.0000
month10      -12.320827 1.9637718 -6.274063  0.0000
month11      -12.650490 2.1922272 -5.770611  0.0000
month12       -9.235813 2.2264505 -4.148223  0.0000
DML:month2     0.018032 0.0083101  2.169894  0.0303
DML:month3     0.003151 0.0068657  0.458989  0.6464
DML:month4     0.002972 0.0074316  0.399848  0.6894
DML:month5    -0.008677 0.0072378 -1.198801  0.2310
DML:month6     0.017620 0.0084180  2.093183  0.0367
DML:month7     0.004647 0.0079929  0.581432  0.5611
DML:month8     0.000500 0.0105569  0.047373  0.9622
DML:month9     0.044242 0.0062130  7.120835  0.0000
DML:month10    0.039495 0.0059673  6.618449  0.0000
DML:month11    0.046671 0.0074640  6.252829  0.0000
DML:month12    0.034099 0.0077022  4.427127  0.0000
```

Selected Model

```
                 Value Std.Error   t-value p-value
(Intercept) -4.780799 1.3347084 -3.581905  0.0004
DML          0.046587 0.0047135  9.883629  0.0000
month2       1.334866 1.6540452  0.807031  0.4199
month3       3.661194 1.5402841  2.376960  0.0177
month4       3.410094 1.5457315  2.206136  0.0277
month5       3.953020 1.5990089  2.472169  0.0137
month6      -0.666428 1.7407663 -0.382836  0.7020
month7       3.669079 1.3886026  2.642281  0.0084
month8       3.241259 1.4361065  2.256977  0.0243
month9      -1.646933 1.4697410 -1.120560  0.2628
month10      0.520224 1.4226828  0.365664  0.7147
month11     -1.742238 1.5319335 -1.137280  0.2558
month12     -0.957199 1.4891726 -0.642773  0.5206
DML:month2  -0.007741 0.0072987 -1.060631  0.2892
DML:month3  -0.020683 0.0058091 -3.560466  0.0004
DML:month4  -0.023433 0.0061142 -3.832590  0.0001
DML:month5  -0.028571 0.0066471 -4.298245  0.0000
DML:month6  -0.012384 0.0075008 -1.651075  0.0991
DML:month7  -0.036971 0.0053041 -6.970336  0.0000
DML:month8  -0.036261 0.0052885 -6.856600  0.0000
DML:month9  -0.002746 0.0056084 -0.489542  0.6246
DML:month10 -0.008962 0.0055068 -1.627426  0.1041
DML:month11  0.007326 0.0061465  1.191932  0.2337
DML:month12  0.007205 0.0061136  1.178596  0.2389
```

Overview of Variance Structures

We covered several possible ways to model heteroskedastic data:

Type	**Formula**	**DF**	`R` **Function**
Fixed	$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \times x_i)$	0	`varFixed()`
Constant	$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$	j-1	`varIdent()`
Power	$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \times \lvert x_i \rvert^{2\delta_j})$	1 or j	`varPower()`
Exponential	$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \times e^{2\delta_j \times x_i})$	1 or j	`varExp()`
Const.+Power	$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \times (\delta_{1j} + \lvert x_i \rvert^{\delta_{2j}})^2)$	1 or 2j	`varConstPower()`
Combination	Variable	Var.	`varComb()`

References

Smith, J.M., Pierce, G.J., Zuur, A.F. & Boyle, P.R. (2005). Seasonal patterns of investment in reproductive and somatic tissues in the squid loligo forbesi. *Aquatic Living Resources*, 18, 341–351.

Zuur et al. 2009 — Chapter 4