Dealing with Heteroskedasticity

Michael Noonan

Biol 520C: Statistical modelling for biological data
1. Heteroskedasticity and the IID Assumption

2. Variance Structures
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:

\[ y_i = X_i \beta + Z_i b_i + \varepsilon_i \]

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

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

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.
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 $x$.

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

More generally, if the variance-covariance matrix of $\varepsilon_i$ 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.
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.
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 \textit{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:

\[ TW_i = \beta_0 + \beta_1 \text{DML}_i + \beta_2 \text{month}_i + \beta_3 \text{DML}_i : \text{month}_i + \varepsilon_i \]
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 = >σ², but different months also have different σ².
Residual spread changes month to month, and in some months increases with DML. So both need to be modelled for a full correction.
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
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 \( \sigma^2 \) 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.
For our squid data, variance increased with DML, so a reasonable formulation to handle the heteroskedasticity would be:
\[
\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \times DML_i)
\]

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

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

anova(mod, mod2)
```

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod</td>
<td>1</td>
<td>3752.084</td>
<td>3867.385</td>
<td>-1851.042</td>
</tr>
<tr>
<td>mod2</td>
<td>2</td>
<td>3620.898</td>
<td>3736.199</td>
<td>-1785.449</td>
</tr>
</tbody>
</table>

AIC shows an improvement, but the residuals still show signs of heterosk.
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 \( \sigma^2 \) and a reference \( \sigma^2 \), so \( j - 1 \) new coefficients are needed to represent the variance function.

\[ \sigma_1^2 = \sigma^2 \times 1, \quad \sigma_2^2 = \sigma^2 \times \theta_1 \quad \ldots \quad \sigma_j^2 = \sigma^2 \times \theta_{j-1} \]
For our squid data, variance differed across months, so we can have the variance change across months: $\varepsilon_{i,\text{month}} \sim \mathcal{N}(0, \sigma^2_{\text{month}})$

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

```r
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.
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 $\delta$ is used for each factor level.
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.

```r
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

```r
anova(mod, mod4)
```

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Test</th>
<th>L.Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod</td>
<td>1</td>
<td>3752.084</td>
<td>3867.385</td>
<td>-1851.042</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mod4</td>
<td>2</td>
<td>3473.019</td>
<td>3592.932</td>
<td>-1710.509</td>
<td>1 vs 2</td>
<td>281.0648</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

AIC shows a big improvement, and the residuals look much better.
Power Variance in \( \mathbb{R} \) cont.

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

\[
\text{mod5} \leftarrow \text{gls}(\text{TW} \sim \text{DML}*\text{month},
\quad \text{weights} = \text{varPower}(\text{form} = \sim \text{DML}|\text{month}),
\quad \text{data} = \text{data})
\]

\text{summary(mod5)}

...  

**Variance function:**
- Structure: Power of variance covariate, different strata
- Formula: \(~DML \mid \text{month}\)
- Parameter estimates:

\[
\begin{align*}
2 & \quad 9 & \quad 12 & \quad 11 & \quad 8 & \quad 10 \\
1.728531 & \quad 1.789503 & \quad 1.733561 & \quad 1.749264 & \quad 1.617694 & \quad 1.789187 \\
5 & \quad 7 & \quad 6 & \quad 4 & \quad 1 & \quad 3 \\
1.746517 & \quad 1.673244 & \quad 1.754488 & \quad 1.711371 & \quad 1.698945 & \quad 1.722742 \\
\end{align*}
\]

...  

\text{anova(mod, mod5)}

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Test</th>
<th>L.Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod</td>
<td>1</td>
<td>3752.084</td>
<td>3867.385</td>
<td>-1851.042</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mod5</td>
<td>2</td>
<td>3407.511</td>
<td>3578.156</td>
<td>-1666.755</td>
<td>1 vs 2</td>
<td>368.5728</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Best AIC yet, and the residuals look pretty good.
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 \( \sigma^2 \) 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, \( \delta \) 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.

```r
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

```r
anova(mod, mod6)
```

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Test L.Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod</td>
<td>1</td>
<td>3752.084</td>
<td>3867.385</td>
<td>-1851.042</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mod6</td>
<td>2</td>
<td>3478.152</td>
<td>3598.066</td>
<td>-1713.076</td>
<td>275.9312</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

AIC shows an improvement, and the residuals look ok.
Again, we can assign different variances for different months via the \texttt{varExp()} function.

\begin{verbatim}
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)
\end{verbatim}

\begin{verbatim}
       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
\end{verbatim}

AIC shows an improvement, and the residuals look good.
The penultimate variance structure we’ll look at is the constant plus power variance structure:

\[ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \times (\delta_1 j + |x_i|\delta_2)^2) \]

- 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 \( \delta_1 \) & \( \delta_2 \) can be grouped by a categorical variable \( j \).
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 \texttt{varConstPower}.

```r
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

```r
anova(mod, mod8)
```

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Test L.Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mod</td>
<td>1</td>
<td>3752.084</td>
<td>3867.385</td>
<td>-1851.042</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mod8</td>
<td>2</td>
<td>3475.019</td>
<td>3599.544</td>
<td>-1710.509</td>
<td>1 vs 2 281.065</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

AIC shows an improvement, and the residuals look ok.
Again, we can assign different variances for different months via the `varConstPower()` function.

```r
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!
The last variance structure we’ll look at is the combined variance structure.

This is an R function (\texttt{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

```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:

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>9</td>
<td>12</td>
<td>11</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>1.0000000</td>
<td>1.4806117</td>
<td>1.0577473</td>
<td>1.1948695</td>
<td>0.5829504</td>
<td>1.4255755</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1.1276263</td>
<td>0.7353367</td>
<td>1.2712401</td>
<td>0.9710625</td>
<td>0.8668377</td>
<td>1.0561607</td>
</tr>
</tbody>
</table>

Structure: Exponential of variance covariate
Formula: ~DML
Parameter estimates:
  `expon`
  0.006711743

```r
anova(mod, mod10)
```

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Test</th>
<th>L.Ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mod</td>
<td>1</td>
<td>25</td>
<td>3752.084</td>
<td>3867.385</td>
<td>-1851.042</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>mod10</td>
<td>2</td>
<td>37</td>
<td>3414.817</td>
<td>3585.463</td>
<td>-1670.409</td>
<td>1 vs 2</td>
<td>361.2663</td>
</tr>
</tbody>
</table>

Original

Combined Variances

Biol 520C: Statistical modelling for biological data
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.

```r
# Calculate AIC values
TABLE <- AIC(mod1, 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

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.215222</td>
<td>1.874068</td>
<td>1.715637</td>
<td>0.0866</td>
</tr>
<tr>
<td>DML</td>
<td>0.021157</td>
<td>0.005558</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>month2</td>
<td>-6.720020</td>
<td>2.328217</td>
<td>-2.886337</td>
<td>0.0440</td>
</tr>
<tr>
<td>month3</td>
<td>-3.926923</td>
<td>2.208195</td>
<td>-1.778340</td>
<td>0.0758</td>
</tr>
<tr>
<td>month4</td>
<td>-4.772045</td>
<td>2.287535</td>
<td>-2.086108</td>
<td>0.0373</td>
</tr>
<tr>
<td>month5</td>
<td>-2.771457</td>
<td>2.265671</td>
<td>-1.223239</td>
<td>0.2216</td>
</tr>
<tr>
<td>month6</td>
<td>-9.598061</td>
<td>2.416310</td>
<td>-3.972196</td>
<td>0.0001</td>
</tr>
<tr>
<td>month7</td>
<td>-7.494959</td>
<td>2.293313</td>
<td>-3.268178</td>
<td>0.0011</td>
</tr>
<tr>
<td>month8</td>
<td>-7.479426</td>
<td>2.931087</td>
<td>-2.551758</td>
<td>0.0109</td>
</tr>
<tr>
<td>month9</td>
<td>-14.963009</td>
<td>2.035313</td>
<td>-7.351698</td>
<td>0.0000</td>
</tr>
<tr>
<td>month10</td>
<td>-12.320827</td>
<td>1.963771</td>
<td>-6.274063</td>
<td>0.0000</td>
</tr>
<tr>
<td>month11</td>
<td>-12.650490</td>
<td>2.192227</td>
<td>-5.770611</td>
<td>0.0000</td>
</tr>
<tr>
<td>month12</td>
<td>-9.235813</td>
<td>2.226450</td>
<td>-4.148223</td>
<td>0.0000</td>
</tr>
<tr>
<td>DML:month2</td>
<td>0.018032</td>
<td>0.008310</td>
<td>0.0303</td>
<td></td>
</tr>
<tr>
<td>DML:month3</td>
<td>0.003151</td>
<td>0.006866</td>
<td>0.458989</td>
<td>0.6464</td>
</tr>
<tr>
<td>DML:month4</td>
<td>0.002972</td>
<td>0.007431</td>
<td>0.399848</td>
<td>0.6894</td>
</tr>
<tr>
<td>DML:month5</td>
<td>-0.008677</td>
<td>0.007237</td>
<td>-1.198801</td>
<td>0.2310</td>
</tr>
<tr>
<td>DML:month6</td>
<td>0.017620</td>
<td>0.008418</td>
<td>0.2093183</td>
<td>0.0367</td>
</tr>
<tr>
<td>DML:month7</td>
<td>0.004647</td>
<td>0.007992</td>
<td>0.581432</td>
<td>0.5611</td>
</tr>
<tr>
<td>DML:month8</td>
<td>0.000500</td>
<td>0.010556</td>
<td>0.047373</td>
<td>0.9622</td>
</tr>
<tr>
<td>DML:month9</td>
<td>0.044242</td>
<td>0.006213</td>
<td>7.120835</td>
<td>0.0000</td>
</tr>
<tr>
<td>DML:month10</td>
<td>0.034945</td>
<td>0.005967</td>
<td>6.618449</td>
<td>0.0000</td>
</tr>
<tr>
<td>DML:month11</td>
<td>0.046671</td>
<td>0.007464</td>
<td>6.252829</td>
<td>0.0000</td>
</tr>
<tr>
<td>DML:month12</td>
<td>0.034099</td>
<td>0.007702</td>
<td>4.427127</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

## Selected Model

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-4.780799</td>
<td>1.334708</td>
<td>-3.581905</td>
<td>0.0004</td>
</tr>
<tr>
<td>DML</td>
<td>0.046587</td>
<td>0.004713</td>
<td>9.883629</td>
<td>0.0000</td>
</tr>
<tr>
<td>month2</td>
<td>1.334866</td>
<td>1.654045</td>
<td>0.807031</td>
<td>0.4199</td>
</tr>
<tr>
<td>month3</td>
<td>3.661194</td>
<td>1.540284</td>
<td>2.376960</td>
<td>0.0177</td>
</tr>
<tr>
<td>month4</td>
<td>3.410094</td>
<td>1.545731</td>
<td>2.206136</td>
<td>0.0277</td>
</tr>
<tr>
<td>month5</td>
<td>3.953020</td>
<td>1.599009</td>
<td>2.472169</td>
<td>0.0137</td>
</tr>
<tr>
<td>month6</td>
<td>-0.666428</td>
<td>1.740766</td>
<td>-0.382836</td>
<td>0.7020</td>
</tr>
<tr>
<td>month7</td>
<td>3.669079</td>
<td>1.386026</td>
<td>2.642281</td>
<td>0.0084</td>
</tr>
<tr>
<td>month8</td>
<td>3.241259</td>
<td>1.436106</td>
<td>2.256977</td>
<td>0.0243</td>
</tr>
<tr>
<td>month9</td>
<td>-1.646933</td>
<td>1.469741</td>
<td>-1.120560</td>
<td>0.2628</td>
</tr>
<tr>
<td>month10</td>
<td>0.520224</td>
<td>1.422682</td>
<td>0.365664</td>
<td>0.7147</td>
</tr>
<tr>
<td>month11</td>
<td>-1.742238</td>
<td>1.531935</td>
<td>-1.137280</td>
<td>0.2558</td>
</tr>
<tr>
<td>month12</td>
<td>-0.957199</td>
<td>1.489172</td>
<td>-0.642773</td>
<td>0.5206</td>
</tr>
<tr>
<td>DML:month2</td>
<td>-0.007741</td>
<td>0.007298</td>
<td>-1.060631</td>
<td>0.2892</td>
</tr>
<tr>
<td>DML:month3</td>
<td>-0.020683</td>
<td>0.005809</td>
<td>-3.560466</td>
<td>0.0004</td>
</tr>
<tr>
<td>DML:month4</td>
<td>-0.023433</td>
<td>0.006142</td>
<td>-3.832590</td>
<td>0.0001</td>
</tr>
<tr>
<td>DML:month5</td>
<td>-0.028571</td>
<td>0.006471</td>
<td>-4.298245</td>
<td>0.0000</td>
</tr>
<tr>
<td>DML:month6</td>
<td>-0.012384</td>
<td>0.007500</td>
<td>-1.651075</td>
<td>0.0991</td>
</tr>
<tr>
<td>DML:month7</td>
<td>-0.036971</td>
<td>0.005304</td>
<td>-6.970336</td>
<td>0.0000</td>
</tr>
<tr>
<td>DML:month8</td>
<td>-0.036261</td>
<td>0.005288</td>
<td>-6.856600</td>
<td>0.0000</td>
</tr>
<tr>
<td>DML:month9</td>
<td>-0.002746</td>
<td>0.005608</td>
<td>-0.489542</td>
<td>0.6246</td>
</tr>
<tr>
<td>DML:month10</td>
<td>-0.008962</td>
<td>0.005506</td>
<td>-1.627426</td>
<td>0.1041</td>
</tr>
<tr>
<td>DML:month11</td>
<td>0.007326</td>
<td>0.006146</td>
<td>1.191932</td>
<td>0.2337</td>
</tr>
<tr>
<td>DML:month12</td>
<td>0.007205</td>
<td>0.006113</td>
<td>1.178596</td>
<td>0.2389</td>
</tr>
</tbody>
</table>
We covered several possible ways to model heteroskedastic data:

<table>
<thead>
<tr>
<th>Type</th>
<th>Formula</th>
<th>DF</th>
<th>R Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed</td>
<td>$\varepsilon_i \sim \mathcal{N}(0, \sigma^2 \times x_i)$</td>
<td>0</td>
<td>varFixed()</td>
</tr>
<tr>
<td>Constant</td>
<td>$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$</td>
<td>j-1</td>
<td>varIdent()</td>
</tr>
<tr>
<td>Power</td>
<td>$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \times</td>
<td>x_i</td>
<td>^{2\delta_j})$</td>
</tr>
<tr>
<td>Exponential</td>
<td>$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \times e^{2\delta_j \times x_i})$</td>
<td>1 or j</td>
<td>varExp()</td>
</tr>
<tr>
<td>Const.+Power</td>
<td>$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \times (\delta_{1j} +</td>
<td>x_i</td>
<td>^{2\delta_j})^2)$</td>
</tr>
<tr>
<td>Combination</td>
<td>Variable</td>
<td>Var.</td>
<td>varComb()</td>
</tr>
</tbody>
</table>
References


Zuur et al. 2009 — Chapter 4