Correlation Structure

Lecture 07.1: Structured Residuals Across Groups

Lauren Sullivan

Module: Linear, Nonlinear, and Mixed Effects Models

Readings

Required for class:

- NA

Optional:

- M. Clark Mixed Models - Extensions for Residual Structure
- phylogenetic mixed effects model

Residual Structure - Constant

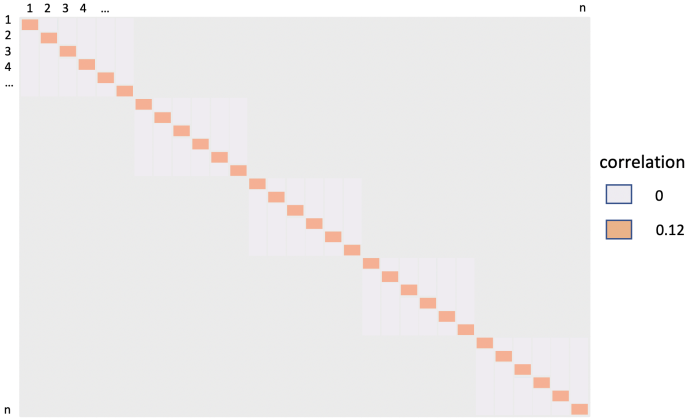
When we have models that assume independence among our data, we assume there is a constant variance across the data and no covariance.

$$\mathbf{y} \sim \mathbf{N}(\mu, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

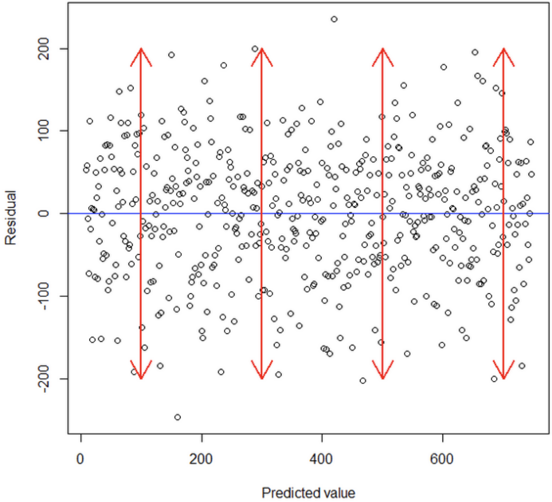
Residual Structure - Constant

Imagine the correlation matrix for an entire dataset. Here, you have variance (σ^2) for each individual along the diagonal (orange) and there is no covariance among individuals (gray).



Residual Structure - Constant

In this case, when you plot your residuals against your predicted values, the variance is equal (as indicated by the red arrows that approximate the variance), and there is no trend.



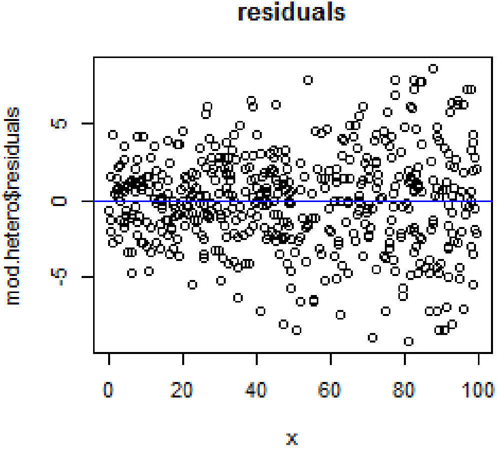
Residual Structure - Varying

We can also relax the assumption of equal variance and estimate each separately. Our covariance matrix (in its simplest form) now looks like this.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_5^2 \end{bmatrix}$$

Residual Structure - Varying

In this case, when you plot your residuals against your predicted values, you see the variance is increasing with increasing x.



Residual Structure - Varying

To make more complicated residual patterns, we need to think about the underlying covariance/correlation. Let's switch to a correlation structure, but still think about the variances as constant or separately estimated. Here, ρ represents the residual correlation among observations.

$$\mathbf{\Sigma} = \sigma_{\mathbf{i}}^{\mathbf{2}} \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_1 & 1 & \rho_5 & \rho_6 & \rho_7 \\ \rho_2 & \rho_5 & 1 & \rho_8 & \rho_9 \\ \rho_3 & \rho_6 & \rho_8 & 1 & \rho_{10} \\ \rho_4 & \rho_7 & \rho_9 & \rho_{10} & 1 \end{bmatrix}$$

This matrix is “symetric” because you can fold it in half along the diagonal and the ρ values are the same.

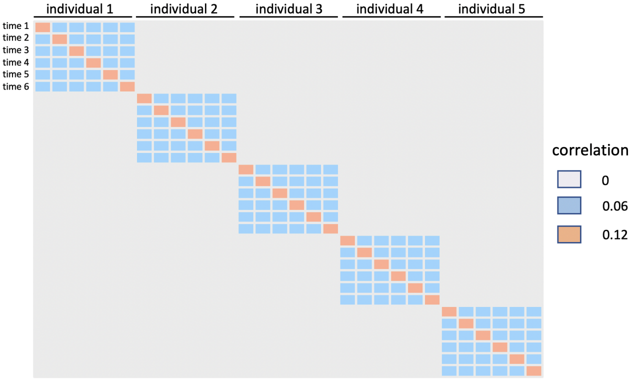
Residual Structure - Varying

So imagine you had 5 measurements and each measurement is correlated with the next one in a sequences, but no others, and that correlation is equal. You might expect the correlation matrix to look something like this.

$$\mathbf{\Sigma} = \sigma_{\mathbf{i}}^{\mathbf{2}} \begin{bmatrix} 1 & 0.06 & 0 & 0 & 0 \\ 0.06 & 1 & 0.06 & 0 & 0 \\ 0 & 0.06 & 1 & 0.06 & 0 \\ 0 & 0 & 0.06 & 1 & 0.06 \\ 0 & 0 & 0 & 0.06 & 1 \end{bmatrix}$$

Residual Structure - Varying

In this case, we are saying that measurements within individuals are correlated by the same ammount (by 0.06 - blue) but there is no covariance across individuals (gray).



Differences in σ_i^2 across time could be multiplied by the covariances that are constant within each individual.

Structured Residuals Example

Let's try to understand this with an example. We have a dataset on individual mystery snails, the biomass consumed for each snail over the course of 6 months, and their sex. We want to find out if male or female snails consume more biomass and if this is dependent on month sampled.

Snail Data

```
## # A tibble: 1,200 x 5
##    individual month biomass sex    month_name
##         <dbl> <dbl>   <dbl> <chr>  <chr>
##  1          1     1    11.5 female month 1
##  2          1     2    10.5 female month 2
##  3          1     3    15   female month 3
##  4          1     4    15   female month 4
##  5          1     5    15   female month 5
##  6          1     6    16.5 female month 6
##  7          2     1    11   male   month 1
##  8          2     2    12.5 male   month 2
##  9          2     3    13   male   month 3
## 10          2     4    13   male   month 4
## # ... with 1,190 more rows
```

Analysis - Mixed Effects Model

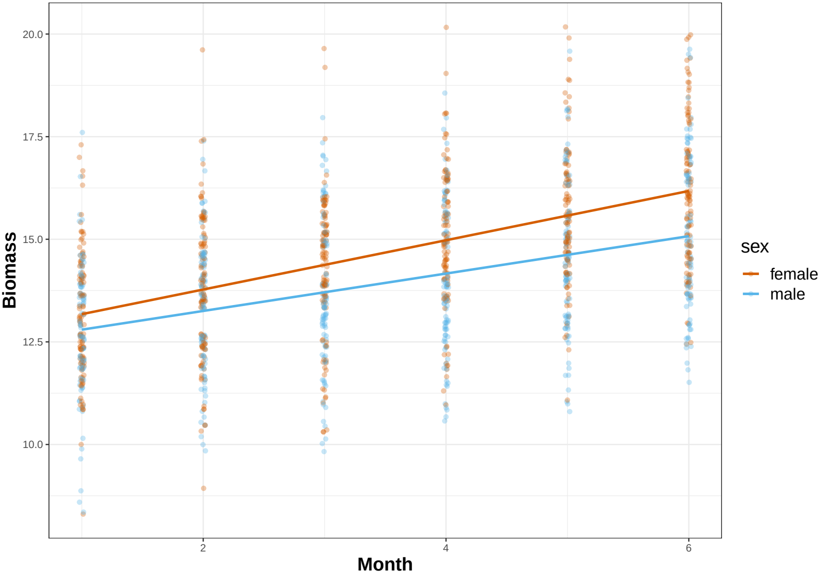
One way to model this is with a mixed effects model.

```r
mod_lme <- lmer(biomass ~ sex + month + (1|individual),
                data = snails, REML = TRUE)
Anova(mod_lme, type = 3)
```

```
## Analysis of Deviance Table (Type III Wald chisquare tests)
##
## Response: biomass
##                Chisq Df Pr(>Chisq)
## (Intercept) 7697.911  1  < 2.2e-16 ***
## sex           15.976  1  6.417e-05 ***
## month        681.021  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

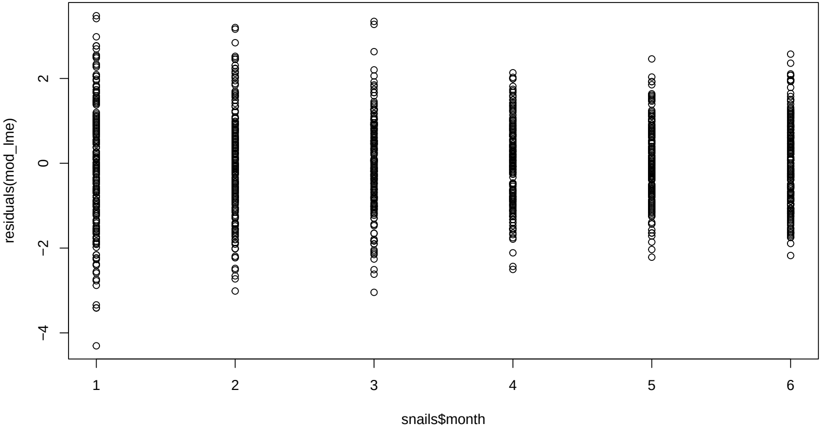
So it seems that male and female snails consume different amounts of biomass, and there is a significant relationship with month and biomass consumed.

Analysis - Random Effects Model



Residual correlation?

But we know that there could be different variance structure through time, as months may differ for these snails.



Here we clearly see that there are differences in our residuals across months.

Residual structure

So there is clearly residual structure despite having equal correlation within individuals. So we will need to specify our σ_i^2's differently.

$$\mathbf{\Sigma} = \sigma_{\mathbf{i}}^{\mathbf{2}} \begin{bmatrix} 1 & 0.06 & 0 & 0 & 0 \\ 0.06 & 1 & 0.06 & 0 & 0 \\ 0 & 0.06 & 1 & 0.06 & 0 \\ 0 & 0 & 0.06 & 1 & 0.06 \\ 0 & 0 & 0 & 0.06 & 1 \end{bmatrix}$$

Analysis - `lme()` with weights

To more appropriately deal with the correlation in time across months, you can parameterize weights in a mixed effects model using `lme()` that has slightly different formatting that `lme4()`.

- The way `lme4()` runs under the hood makes it very difficult to include weights, so when you want to account for a correlation structure, use `lme()`. This will allow the model to estimate a different variance for each month.

Check out this page for documentation on how to specify the `weights` argument, and this page for the different type of classes of `weights`.

- In our case we will want to use `varIdent` because this allows for different variances according to the level of a classification factor (or here, month).

Analysis - `lme()` with weights

Here, we are going to specify our `weights = varIdent()` using this page. We will specify our starting variance (`v`) as 1, and the grouping factor (`g`) as the month.

- Again notice the slightly different form for a mixed effects model in `lme()` vs what we used last time with `lme4()`.

```r
mod_corr_month <- lme(biomass ~ sex + month,
                      data = snails,
                      random = ~1|individual,
                      weights = varIdent(form = ~1|month))
```

Analysis - Results

```
## Linear mixed-effects model fit by REML
##   Data: snails
##        AIC      BIC    logLik
##   4116.856 4167.732 -2048.428
##
## Random effects:
##  Formula: ~1 | individual
##         (Intercept) Residual
## StdDev:    1.467741 1.858308
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | month
##  Parameter estimates:
##         1         2         3         4         5         6
## 1.0000000 0.8268408 0.6285354 0.4306586 0.3477960 0.4321086
## Fixed effects:  biomass ~ sex + month
##                 Value  Std.Error  DF  t-value p-value
## (Intercept) 12.887113 0.17290706 999 74.53202       0
## sexmale     -0.892425 0.21465607 198 -4.15746       0
## month        0.530705 0.02016551 999 26.31746       0
##  Correlation:
##         (Intr) sexmal
## sexmale -0.590
## month   -0.518  0.000
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.76463094 -0.63616250  0.02437261  0.62482431  3.07404044
##
## Number of Observations: 1200
## Number of Groups: 200
```

Analysis - Results

So here you again see that males consume less biomass than females, and that there is a positive increase in biomass consumed through time. However you also see that there are different parameters estimated for each month.

Note: With `lme()`, you will some times need to look up your p-values with a table (remember those?). Here, we would need a table of t-values, and you want the one-tailed row.

For example, in our case for sex you have: t-value = -4.15, and DF = 198.

- So your p-value < 0.0005

Residual estimates

With this formulation, you get your residual values in standard deviation instead of variance, and as you can see your variance decreases, but the actual values are not provided.

```
summary(mod_corr_month$modelStruct)
```

```
## Random effects:
##  Formula: ~1 | individual
##         (Intercept) Residual
## StdDev:   0.7898267        1
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | month
##  Parameter estimates:
##         1         2         3         4         5         6
## 1.0000000 0.8268408 0.6285354 0.4306586 0.3477960 0.4321086
```

Residual estimates

To get actual estimates of your residuals, you will want to scale them and square them to get them on the variance scale.

```
(c(1.0000000, coef(mod_corr_month$modelStruct$varStruct, unconstrained=F))*mod_corr_month$sigma)^2
```

```
##                   2         3         4         5         6
## 3.4533094 2.3609095 1.3642531 0.6404744 0.4177195 0.6447944
```