Смешанные линейные модели (вложенные случайные факторы)

Линейные модели...

Марина Варфоломеева, Вадим Хайтов
Осень 2022

Вы узнаете

- Что такое вложенные случайные факторы и в каких случаях они применяются

Вы сможете

- Объяснить, что такое вложенные случайные факторы
- Привести примеры иерархических случайных факторов
- Вычислить коэффициент внутриклассовой корреляции для случая с двумя вложенными случайными факторами
- Подобрать смешанную линейную модель со вложенными случайными факторами

Смешанные модели со вложенными случайными факторами

Вложенные факторы (Nested effects)

Вложенные факторы (Nested effects)

Факторы образуют иерархическую последовательность вложенности

- лес –> дерево в лесу –> ветка на дереве –> наблюдение (личинки насекомых)

Вложенные факторы (Nested effects)

Факторы образуют иерархическую последовательность вложенности

- лес –> дерево в лесу –> ветка на дереве –> наблюдение (личинки насекомых)

Внутри каждого уровня главного фактора будут разные (нестрого сопоставимые) уровни вложенного фактора

Деревья, с которых собирали личинок, будут разные в разных лесах (разные экземпляры).

Вложенные факторы (Nested effects)

Факторы образуют иерархическую последовательность вложенности

- лес –> дерево в лесу –> ветка на дереве –> наблюдение (личинки насекомых)

Внутри каждого уровня главного фактора будут разные (нестрого сопоставимые) уровни вложенного фактора

Деревья, с которых собирали личинок, будут разные в разных лесах (разные экземпляры).

Уровни вложенных факторов описывают иерархию взаимного сходства наблюдений

Личинки с разных деревьев из одного леса имеют право быть похожими друг на друга больше, чем на личинок из другого леса
Личинки на одном дереве имеют право быть похожими друг на друга больше, чем на личинок с другого дерева
И т.п.

Другие примеры вложенных факторов

Знакомство с данными

Есть ли пропущенные значения?

```
sum(is.na(graz))
```

```
[1] 0
```

Сколько участков было в каждом парке в каждый год?

```
with(graz, table(Park, year))
```

```
    year
Park 2004 2005 2006 2007 2008 2009 2010 2011
  MT    6   10   10   10   10   10   10   10
  PR    6    6    6    6    6    6    6    6
  SU    0    9    9    9    9    9    9    9
  VC   10   10   10   10   11   11   11   11
```

Как закодированы переменные?

```
str(graz)
```

```
tibble [271 × 18] (S3: tbl_df/tbl/data.frame)
 $ plotID     : chr [1:271] "MT1" "MT2" "MT3" "MT4" ...
 $ Plot       : num [1:271] 1 2 3 4 5 6 4 5 6 7 ...
 $ Park       : chr [1:271] "MT" "MT" "MT" "MT" ...
 $ year       : num [1:271] 2004 2004 2004 2004 2004 ...
 $ graze      : num [1:271] 0 0 0 1 1 1 1 1 1 1 ...
 $ Aspect     : num [1:271] 146 250 262 190 274 ...
 $ AspectCat  : chr [1:271] "S" "S" "S" "S" ...
 $ heatloadrel: num [1:271] 0.03 1.35 2.06 0.31 1.83 0.95 0.03 2.06 0.7 1.22 ...
 $ slope      : num [1:271] 37.8 41.1 35.4 28 58.9 ...
 $ nativecov  : num [1:271] 0 0.36 1.43 0 1.07 ...
 $ litt       : num [1:271] 28.21 31.43 11.07 8.93 18.57 ...
 $ bare       : num [1:271] 0 0.357 9.286 4.643 7.857 ...
 $ height     : num [1:271] 29.2 26.2 23.2 11.9 19.3 ...
 $ htstdev    : num [1:271] 16.14 15.41 15.47 8.25 10.45 ...
 $ cov        : num [1:271] 0.553 0.587 0.666 0.695 0.541 ...
 $ GRSP       : num [1:271] 0 0 0 0 1 0 1 1 0 0 ...
 $ HOLA       : num [1:271] 0 0 0 1 1 1 1 1 0 0 ...
 $ WEME       : num [1:271] 0 1 0 1 1 0 0 1 0 1 ...
```

Наводим порядок

Сделаем факторами переменные, которые понадобятся для модели

```
graz$graze_f <- factor(graz$graze)
graz$AspectCat <- factor(graz$AspectCat)
graz$year_f <- factor(graz$year)
```

Извлечем корень из обилия местных видов

```
graz$nativecov_sq <- sqrt(graz$nativecov)
```

Модель

Вспомним главный вопрос исследования и подберем модель

Как в разные годы высота растительного покрова зависит от выпаса скота, экспозиции склона и проективного покрытия местных растений?

Модель

Вспомним главный вопрос исследования и подберем модель

Как в разные годы высота растительного покрова зависит от выпаса скота, экспозиции склона и проективного покрытия местных растений?

Нам нужно учесть, что в разные годы из-за кучи разных причин высота растений может различаться

Кроме того, нужно учесть, что в разных парках и на разных участках растения будут расти сходным образом в разные годы. У нас есть иерархические факторы парк и участок в парке

Подбираем модель методом максимального правдоподобия, т.к. она нам понадобится, чтобы проверить значимость фиксированных эффектов при помощи теста отношения правдоподобий.

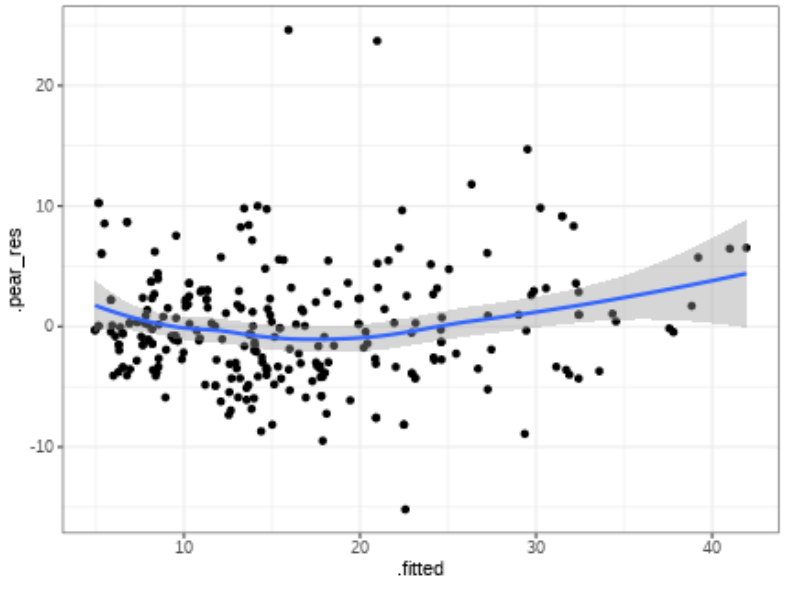
```
library(lme4)
ML1 <- lmer(height ~ graze_f*AspectCat + year_f +
              nativecov_sq + slope + (1|Park/plotID),
          data = graz, REML = FALSE)
```

Анализ остатков

```
# Данные для анализа остатков
ML1_diag <- data.frame(
  graz,
  .pear_res = residuals(ML1, type = "pearson"),
  .fitted = fitted(ML1, type = "response"))
```

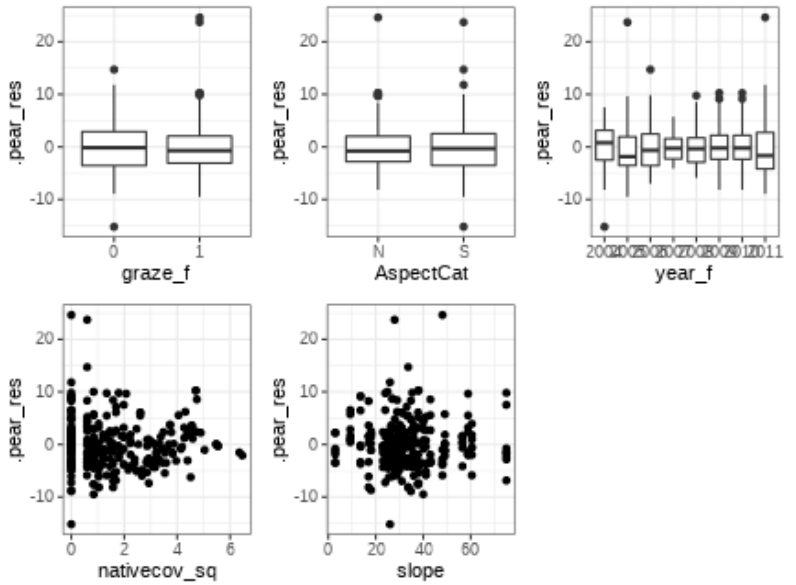
График остатков

```
library(ggplot2); library(cowplot); theme_set(theme_bw())
gg_res <- ggplot(data = ML1_diag, aes(y = .pear_res))
gg_res + geom_point(aes(x = .fitted)) +
  geom_smooth(aes(x = .fitted))
```



Графики остатков от переменных в модели

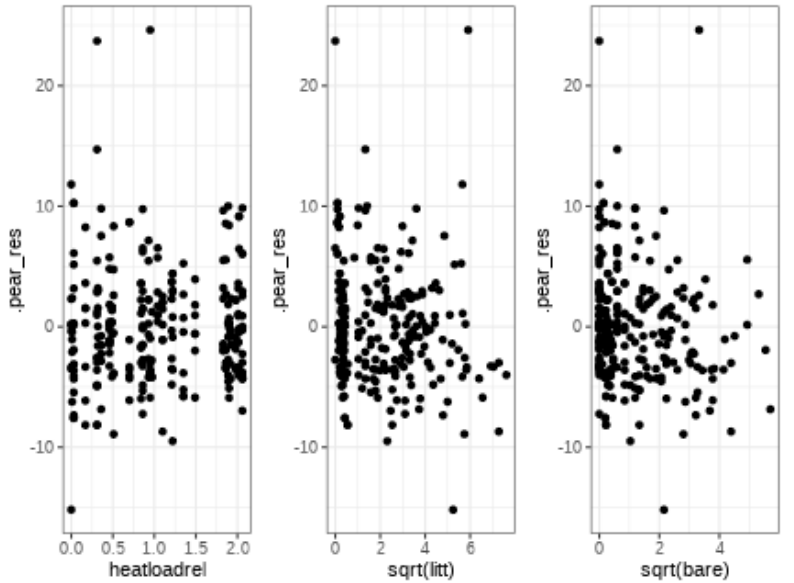
```
plot_grid(gg_res + geom_boxplot(aes(x = graze_f)),
gg_res + geom_boxplot(aes(x = AspectCat)),
gg_res + geom_boxplot(aes(x = year_f)),
gg_res + geom_point(aes(x = nativecov_sq)),
gg_res + geom_point(aes(x = slope)),
ncol = 3)
```



- Паттерн на графике `nativecov_sq`. Возможно, здесь нужно использовать GAMM.

Графики остатков от переменных не в модели

```
plot_grid(
  gg_res + geom_point(aes(x = heatloadrel)),
  gg_res + geom_point(aes(x = sqrt(litt))),
  gg_res + geom_point(aes(x = sqrt(bare))),
  ncol = 3)
```



- Паттерн на графике `heatloadrel`
- Возможно, есть тренд на графике `sqrt(litt)`

Тесты отношения правдоподобий для полной модели

Модель `ML1` была подобрана при помощи ML, поэтому можно применять тесты отношения правдоподобий прямо к ней

```
drop1(ML1, test = 'Chi')
```

```
Single term deletions

Model:
height ~ graze_f * AspectCat + year_f + nativecov_sq + slope +
    (1 | Park/plotID)
                 npar  AIC   LRT Pr(Chi)
<none>               1729
year_f              7 1820 104.4  <2e-16 ***
nativecov_sq        1 1729   1.5  0.2240
slope               1 1728   0.4  0.5025
graze_f:AspectCat   1 1736   9.0  0.0028 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Высота растительного покрова:

- на склонах разной экспозиции по-разному зависит от выпаса скота (достоверное взаимодействие)
- различается в разные годы
- не зависит от покрытия местных растений и крутизны склона

Задание 1

Рассчитайте внутриклассовую корреляцию

- Для наблюдений на одном и том же участке
- Для наблюдений в одном и том же парке

Внутриклассовая корреляция

Для расчета внутриклассовой корреляции нужна модель, подобранная при помощи REML

```
REML1 <- lmer(height ~ graze_f*AspectCat + year_f +
                nativecov_sq + slope + (1|Park/plotID),
           data = graz, REML = TRUE)
```

```
summary(REML1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: height ~ graze_f * AspectCat + year_f + nativecov_sq + slope +
    (1 | Park/plotID)
   Data: graz

REML criterion at convergence: 1678

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.924 -0.637 -0.082  0.433  4.796

Random effects:
 Groups      Name        Variance Std.Dev.
 plotID:Park (Intercept) 11.36    3.37
 Park        (Intercept)  2.48    1.57
 Residual                26.35    5.13
Number of obs: 271, groups:  plotID:Park, 36; Park, 4

Fixed effects:
                   Estimate Std. Error t value
(Intercept)         18.4491     3.0923    5.97
```

Внутриклассовая корреляция

Для наблюдений на одном и том же участке $\sigma^2_{plotID}/(\sigma^2_{plotID} + \sigma^2_{Park} + \sigma^2)$

```
3.370^2 / (1.574^2 + 3.370^2 + 5.133^2)
```

```
[1] 0.2826
```

Для наблюдений в одном и том же парке $\sigma^2_{Park}/(\sigma^2_{plotID} + \sigma^2_{Park} + \sigma^2)$

```
1.574^2 / (1.574^2 + 3.370^2 + 5.133^2)
```

```
[1] 0.06166
```

```
В результатах summary(REML1)
Random effects:
 Groups      Name        Variance Std.Dev.
 plotID:Park (Intercept) 11.358   3.370
 Park        (Intercept)  2.478   1.574
 Residual                26.351   5.133
Number of obs: 271, groups:  plotID:Park, 36; Park, 4
```

- Значения высоты травяного покрова похожи внутри участка. Сходство наблюдений внутри одного парка слабее.

Результаты полной модели

```
summary(REML1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: height ~ graze_f * AspectCat + year_f + nativecov_sq + slope +
    (1 | Park/plotID)
   Data: graz

REML criterion at convergence: 1678

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.924 -0.637 -0.082  0.433  4.796

Random effects:
 Groups      Name        Variance Std.Dev.
 plotID:Park (Intercept) 11.36    3.37
 Park        (Intercept)  2.48    1.57
 Residual                26.35    5.13
Number of obs: 271, groups:  plotID:Park, 36; Park, 4

Fixed effects:
                   Estimate Std. Error t value
(Intercept)         18.4491     3.0923    5.97
graze_f1            -5.2543     2.7131   -1.94
AspectCatS           8.5716     2.7556    3.11
year_f2005           6.8075     1.4239    4.78
year_f2006           4.0797     1.4162    2.88
```

Данные для графика предсказаний фиксированной части модели

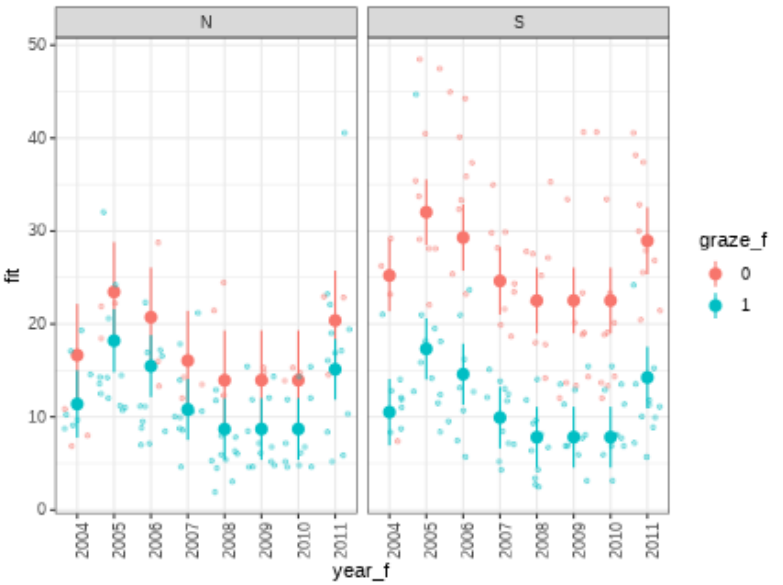
Используем для визуализации модель, подобранную при помощи REML

```
# Исходные данные
NewData_REML1 <- expand.grid(graze_f = levels(graz$graze_f),
            AspectCat = levels(graz$AspectCat),
            year_f = levels(graz$year_f))
NewData_REML1$nativecov_sq <- mean(graz$nativecov_sq)
NewData_REML1$slope <- mean(graz$slope)

# Предсказанные значения при помощи матриц
X <- model.matrix(~ graze_f * AspectCat + year_f + nativecov_sq + slope,
                  data = NewData_REML1)
betas = fixef(REML1)
NewData_REML1$fit <- X %*% betas

# Стандартные ошибки и дов. интервалы
NewData_REML1$se <- sqrt( diag(X %*% vcov(REML1) %*% t(X)) )
NewData_REML1$lwr <- NewData_REML1$fit - 2 * NewData_REML1$se
NewData_REML1$upr <- NewData_REML1$fit + 2 * NewData_REML1$se
```

График предсказаний фиксированной части модели

На южных склонах высота травы выше там, где не пасут скот, а на северных нет.
(Строго говоря, нужен еще пост хок тест, чтобы это утверждать.)

```
ggplot(data = NewData_REML1, aes(x = year_f, y = fit, colour = graze_f)) +
  geom_pointrange(aes(ymin = lwr, ymax = upr)) +
  facet_wrap(~ AspectCat) +
  geom_jitter(data = graz, aes(y = height), alpha = 0.35, size = 1) +
  theme(axis.text.x = element_text(angle = 90))
```



Вариант решения с подбором оптимальной модели (самостоятельно)

Задание 2

Оптимизируйте модель с предыдущего шага

Сделайте анализ остатков

Опишите и визуализируйте финальную модель

Решение: Подбор оптимальной модели (1)

Для подбора оптимальной модели воспользуемся тестами отношения правдоподобий. Для него нужно использовать модели, подобранные при помощи ML

```
drop1(ML1, test = "Chi")
```

```
Single term deletions

Model:
height ~ graze_f * AspectCat + year_f + nativecov_sq + slope +
    (1 | Park/plotID)
                 npar  AIC   LRT Pr(Chi)
<none>                1729
year_f              7 1820 104.4  <2e-16 ***
nativecov_sq        1 1729   1.5  0.2240
slope               1 1728   0.4  0.5025
graze_f:AspectCat   1 1736   9.0  0.0028 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Решение: Подбор оптимальной модели (2)

```
ML1.1 <- update(ML1, .~.-slope)
drop1(ML1.1, test = "Chi")
```

```
Single term deletions

Model:
height ~ graze_f + AspectCat + year_f + nativecov_sq + (1 | Park/plotID) +
    graze_f:AspectCat
                 npar  AIC   LRT Pr(Chi)
<none>                1728
year_f              7 1818 104.3  <2e-16 ***
nativecov_sq        1 1728   1.7  0.1948
graze_f:AspectCat   1 1734   8.5  0.0035 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Решение: Подбор оптимальной модели (3)

```
ML1.2 <- update(ML1.1, .~.-nativecov_sq)
drop1(ML1.2, test = "Chi")
```

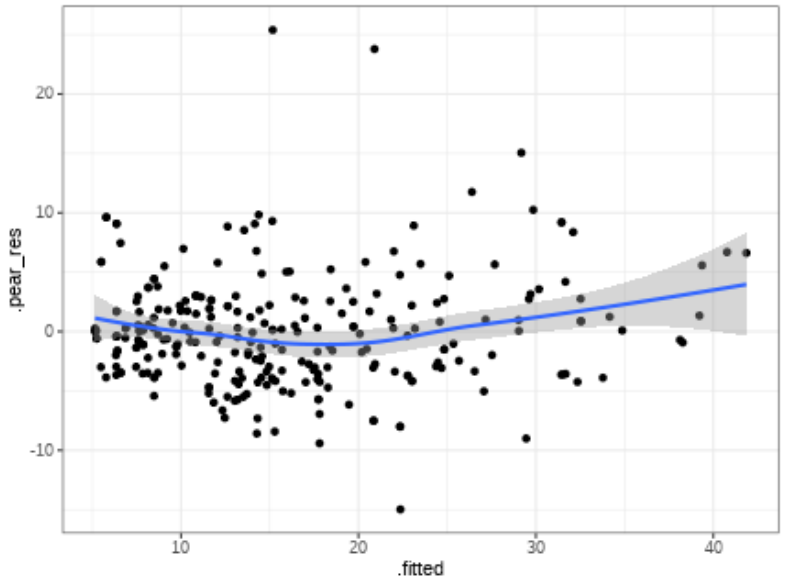
```
Single term deletions

Model:
height ~ graze_f + AspectCat + year_f + (1 | Park/plotID) + graze_f:AspectCat
                  npar  AIC   LRT Pr(Chi)
<none>                 1728
year_f               7 1819 105.2  <2e-16 ***
graze_f:AspectCat    1 1734   8.1  0.0045 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Решение: Анализ остатков

```
# Данные для анализа остатков
ML1.2_diag <- data.frame(
  graz,
  .pear_res = residuals(ML1.2, type = "pearson"),
  .fitted = fitted(ML1.2, type = "response"))
```

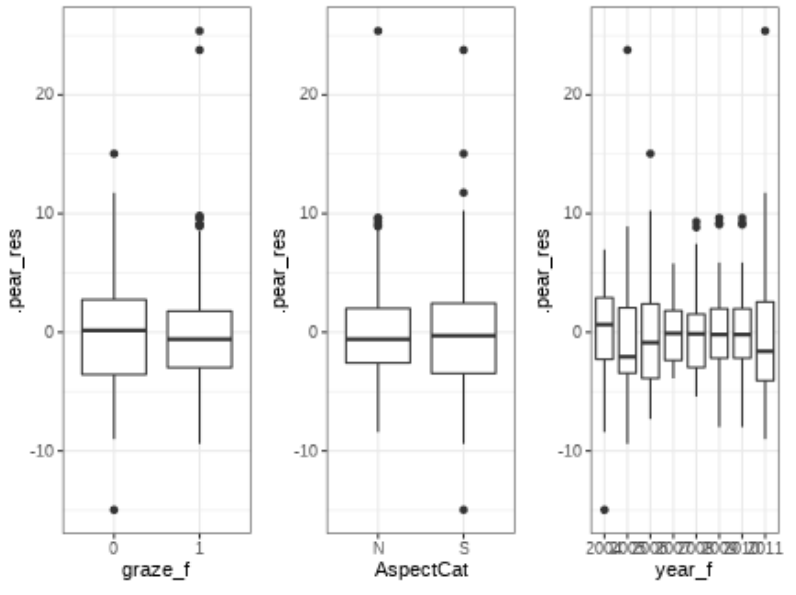
Решение: График остатков

```
gg_res <- ggplot(data = ML1.2_diag, aes(y = .pear_res))
gg_res + geom_point(aes(x = .fitted)) +
  geom_smooth(aes(x = .fitted))
```



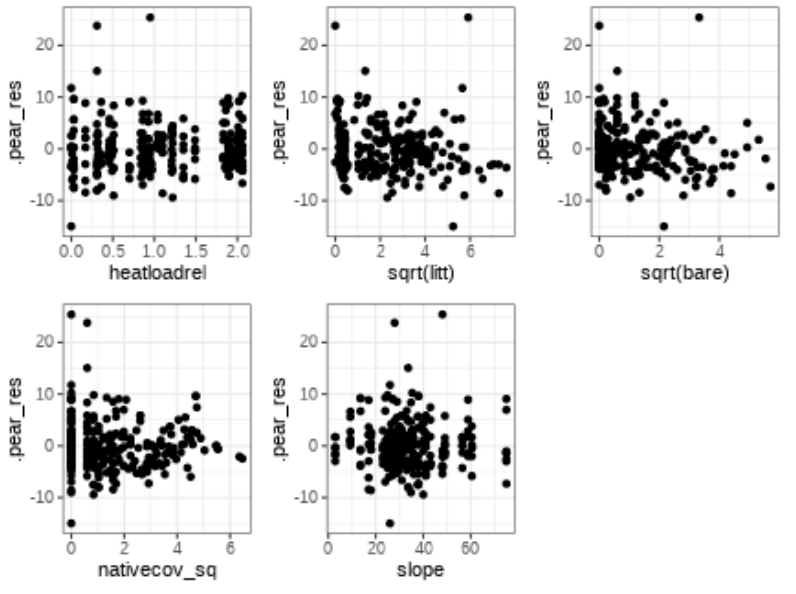
Решение: Графики остатков от переменных в модели

```
plot_grid(gg_res + geom_boxplot(aes(x = graze_f)),
gg_res + geom_boxplot(aes(x = AspectCat)),
gg_res + geom_boxplot(aes(x = year_f)),
ncol = 3)
```



Решение: Графики остатков от переменных не в модели

```
plot_grid(
  gg_res + geom_point(aes(x = heatloadrel)),
  gg_res + geom_point(aes(x = sqrt(litt))),
  gg_res + geom_point(aes(x = sqrt(bare))),
  gg_res + geom_point(aes(x = nativecov_sq)),
  gg_res + geom_point(aes(x = slope)),
  ncol = 3)
```



- Паттерн на графике `heatloadrel`, `nativecov_sq`

Решение: Тестируем влияние факторов в финальной модели

Для тестов отношения правдоподобий используем финальную модель, подобранную при помощи ML

```
drop1(ML1.2, test = 'Chi')
```

```
Single term deletions

Model:
height ~ graze_f + AspectCat + year_f + (1 | Park/plotID) + graze_f:AspectCat
                  npar  AIC   LRT Pr(Chi)
<none>                 1728
year_f               7 1819 105.2  <2e-16 ***
graze_f:AspectCat    1 1734   8.1  0.0045 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Высота растительного покрова:

- на склонах разной экспозиции по-разному зависит от выпаса скота (достоверное взаимодействие)
- различается в разные годы
- не зависит от покрытия местных растений и крутизны склона

Решение: Описываем результаты

Для описания результатов используем модель, подобранную при помощи REML, т.к. он дает более точные оценки случайных эффектов

```
REML1.2 <- update(ML1.2, REML = TRUE)
```

Решение: Результаты

```
summary(REML1.2)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: height ~ graze_f + AspectCat + year_f + (1 | Park/plotID) + graze_f:AspectCat
   Data: graz

REML criterion at convergence: 1676

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.917 -0.631 -0.079  0.426  4.985

Random effects:
 Groups      Name        Variance Std.Dev.
 plotID:Park (Intercept) 13.76    3.71
 Park        (Intercept)  1.38    1.18
 Residual                25.92    5.09
Number of obs: 271, groups:  plotID:Park, 36; Park, 4

Fixed effects:
                   Estimate Std. Error t value
(Intercept)          16.408      2.777    5.91
graze_f1             -5.678      2.760   -2.06
AspectCatS            9.441      2.875    3.28
year_f2005            6.626      1.404    4.72
year_f2006            4.131      1.404    2.94
year_f2007           -0.368      1.404   -0.26
```

Решение: Внутриклассовая корреляция

Для расчета нужна модель, подобранная при помощи REML

```
Random effects:
 Groups      Name        Variance Std.Dev.
 plotID:Park (Intercept) 13.761   3.710
 Park        (Intercept)  1.384   1.177
 Residual                25.916   5.091
Number of obs: 271, groups:  plotID:Park, 36; Park, 4
```

Для наблюдений на одном и том же участке $\sigma^2_{plotID}/(\sigma^2_{plotID} + \sigma^2_{Park} + \sigma^2)$

```
3.710^2 / (1.177^2 + 3.710^2 + 5.091^2)
```

```
[1] 0.3352
```

Для наблюдений в одном и том же парке $\sigma^2_{Park}/(\sigma^2_{plotID} + \sigma^2_{Park} + \sigma^2)$

```
1.177^2 / (1.177^2 + 3.7010^2 + 5.091^2)
```

```
[1] 0.03379
```

- Значения высоты травяного покрова похожи внутри участка. Сходство наблюдений внутри одного парка слабее.

Решение: Данные для графика предсказаний фиксированной части модели

Используем для визуализации модель, подобранную при помощи REML

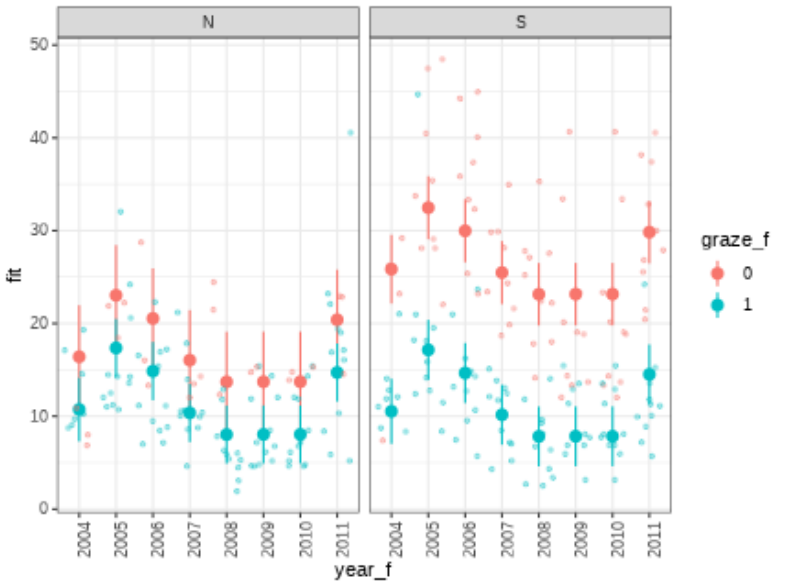
```
# Исходные данные
NewData_REML1.2 <- expand.grid(graze_f = levels(graz$graze_f),
            AspectCat = levels(graz$AspectCat),
            year_f = levels(graz$year_f))
NewData_REML1.2$nativecov_sq <- mean(graz$nativecov_sq)
NewData_REML1.2$slope <- mean(graz$slope)

# Предсказанные значения при помощи матриц
X <- model.matrix(~ graze_f * AspectCat + year_f, data = NewData_REML1.2)
betas = fixef(REML1.2)
NewData_REML1.2$fit <- X %*% betas

# Стандартные ошибки и дов. интервалы
NewData_REML1.2$se <- sqrt( diag(X %*% vcov(REML1.2) %*% t(X)) )
NewData_REML1.2$lwr <- NewData_REML1.2$fit - 2 * NewData_REML1.2$se
NewData_REML1.2$upr <- NewData_REML1.2$fit + 2 * NewData_REML1.2$se
```

Решение: График предсказаний фиксированной части модели

На южных склонах высота травы выше там, где не пасут скот, а на северных нет. (Строго говоря, нужен еще пост хок тест, чтобы это утверждать)

График похож на предыдущий, т.к. удаленные факторы и так не влияли.

```
ggplot(data = NewData_REML1.2, aes(x = year_f, y = fit, colour = graze_f)) +
  geom_pointrange(aes(ymin = lwr, ymax = upr)) +
  facet_wrap(~ AspectCat) +
  geom_jitter(data = graz, aes(y = height), alpha = 0.35, size = 1) +
  theme(axis.text.x = element_text(angle = 90))
```



Take-home messages

- Случайные факторы в смешанных моделях могут быть вложены друг в друга
- Есть два способа подбора коэффициентов в смешанных моделях: ML и REML. Для разных этапов анализа важно, каким именно способом подобрана модель.

Дополнительные ресурсы

- Crawley, M.J. (2007). The R Book (Wiley).
- **Faraway, J. J. (2017). Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models (Vol. 124). CRC press.**
- Zuur, A. F., Hilbe, J., & Ieno, E. N. (2013). A Beginner's Guide to GLM and GLMM with R: A Frequentist and Bayesian Perspective for Ecologists. Highland Statistics.
- **Zuur, A.F., Ieno, E.N., Walker, N., Saveliev, A.A., and Smith, G.M. (2009). Mixed Effects Models and Extensions in Ecology With R (Springer)**
- Pinheiro, J., Bates, D. (2000). Mixed-Effects Models in S and S-PLUS. Springer