Смешанные линейные модели (случайный интерсепт и случайный угол наклона)

Линейные модели...

Марина Варфоломеева

Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ

Вы узнаете

- Что такое смешаные модели и когда они применяются
- Что такое фиксированные и случайные факторы

Вы сможете

- Рассказать чем фиксированные факторы отличаются от случайных
- Привести примеры факторов, которые могут быть фиксированными или случайными в зависимости от задачи исследования
- Рассказать, что оценивает коэффициент внутриклассовой корреляции и вычислить его для случая с одним случайным фактором
- Подобрать смешаную линейную модель со случайным отрезком и случайным углом наклона в R при помощи методов максимального правдоподобия

“Многоуровневые” данные

Независимость наблюдений

Обычные линейные модели предполагают, что наблюдения должны быть независимы друг от друга.

Но так происходит совсем не всегда.

Многоуровневые (multilevel), сгруппированные (clustered) данные

Иногда наблюдения бывают сходны по каким-то признакам:

- измерения в разные периоды времени
  - измерения, сделанные в химической лаборатории в разные дни
- измерения в разных участках пространства
  - урожай на участках одного поля
  - детали, произведенные на одном из нескольких аналогичных станков
- повторные измерения на одних и тех же субъектах
  - измерения до и после какого-то воздействия
- измерения на разных субъектах, которые сами объединены в группы
  - ученики в классах, классы в школах, школы в районах, районы в городах и т.п.

Внутригрупповые корреляции

Детали, произведенные на одном станке будут более похожи, чем детали, сделанные на разных.

Аналогично, у учеников из одного класса будет более похожий уровень подготовки к какому-нибудь предмету, чем у учеников из разных классов.

Таким образом, можно сказать, что есть корреляции значений внутри групп.

Последствия внутригрупповых корреляций для анализа

Игнорировать такую группирующую структуру данных нельзя – можно ошибиться с выводами.

Моделировать группирующие факторы обычными методами тоже нельзя – придется подбирать очень много параметров.

Решение – случайные факторы. О том, что это такое и чем они принципиально отличаются от фиксированных, мы поговорим позже. А сейчас давайте на примере убедимся в том, что без случайных факторов бывает сложно справиться с анализом.

Пример – недосып и время реакции

В ночь перед нулевым днем всем испытуемым давали поспать нормальное время, а в следующие 9 ночей — только по 3 часа. Каждый день измеряли время реакции в серии тестов.

Как время реакции людей зависит от бессонницы?

Данные: Belenky et al., 2003
Источник: пакет `lme4`

Открываем данные

- Reaction — среднее время реакции в серии тестов в день наблюдения, мс
- Days — число дней депривации сна
- Subject — номер испытуемого

```
library(lme4)
data(sleepstudy)

sl <- sleepstudy
str(sl)
```

```
# 'data.frame': 180 obs. of  3 variables:
#  $ Reaction: num  250 259 251 321 357 ...
#  $ Days    : num  0 1 2 3 4 5 6 7 8 9 ...
#  $ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Знакомимся с данными

```r
# Есть ли пропущенные значения?
colSums(is.na(sl))
```

```
# Reaction     Days  Subject
#        0        0        0
```

```r
# Сколько субъектов?
length(unique(sl$Subject))
```

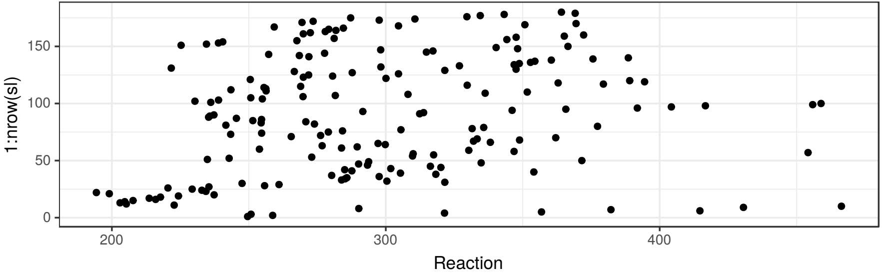
```
# [1] 18
```

```r
# Сколько наблюдений для каждого субъекта?
table(sl$Subject)
```

```
#
# 308 309 310 330 331 332 333 334 335 337 349 350 351 352 369 370 371
#  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10
# 372
#  10
```

Есть ли выбросы?

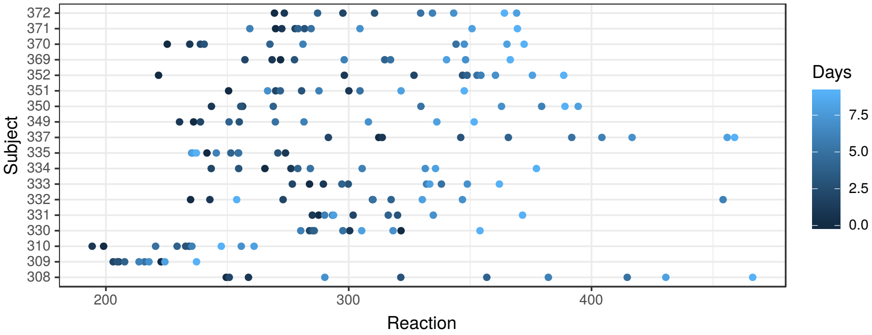
```r
library(ggplot2)
theme_set(theme_bw())

ggplot(sl, aes(x = Reaction, y = 1:nrow(sl))) +
  geom_point()
```



Мы пока еще не учли информацию о субъектах…

Как меняется время реакции разных людей?

```
ggplot(sl, aes(x = Reaction, y = Subject, colour = Days)) +
  geom_point()
```



У разных людей разное время реакции.

Межиндивидуальную изменчивость нельзя игнорировать.

Что делать с разными субъектами?

Что делать с разными субъектами?

The Good — подбираем смешанную модель, в которой есть фиксированный фактор `Days` и случайный фактор `Subject`, который опишет межиндивидуальную изменчивость.

Что делать с разными субъектами?

The Good — подбираем смешанную модель, в которой есть фиксированный фактор `Days` и случайный фактор `Subject`, который опишет межиндивидуальную изменчивость.

The Bad — игнорируем структуру данных, подбираем модель с единственным фиксированным фактором `Days`. (Не учитываем группирующий фактор `Subject`). Неправильный вариант.

Что делать с разными субъектами?

The Good — подбираем смешанную модель, в которой есть фиксированный фактор `Days` и случайный фактор `Subject`, который опишет межиндивидуальную изменчивость.

The Bad — игнорируем структуру данных, подбираем модель с единственным фиксированным фактором `Days`. (Не учитываем группирующий фактор `Subject`). Неправильный вариант.

The Ugly — подбираем модель с двумя фиксированными факторами: `Days` и `Subject`. (Группирующий фактор `Subject` опишет межиндивидуальную изменчивость как обычный фиксированный фактор).

Плохое решение: не учитываем группирующий фактор

$Reaction_i = \beta_0 + \beta_1 Days_i + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma)$

Плохое решение: не учитываем группирующий фактор

$Reaction_i = \beta_0 + \beta_1 Days_i + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma)$

В матричном виде это можно записать так:

$$\begin{pmatrix} Reaction_1 \\ Reaction_2 \\ \vdots \\ Reaction_{180} \end{pmatrix} = \begin{pmatrix} 1 & Days_1 \\ 1 & Days_2 \\ \vdots & \\ 1 & Days_{180} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{180} \end{pmatrix}$$

что можно сокращенно записать так:

$\mathbf{Reaction} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Плохое решение: не учитываем группирующий фактор

```
W1 <- glm(Reaction ~ Days, data = sl)
summary(W1)
```

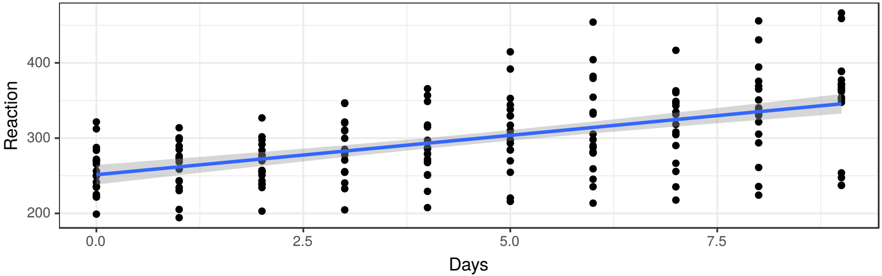
```
#
# Call:
# glm(formula = Reaction ~ Days, data = sl)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -110.848   -27.483     1.546    26.142   139.953
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  251.405      6.610  38.033  < 2e-16 ***
# Days          10.467      1.238   8.454 9.89e-15 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for gaussian family taken to be 2276.694)
#
#     Null deviance: 567954  on 179  degrees of freedom
# Residual deviance: 405252  on 178  degrees of freedom
# AIC: 1906.3
#
# Number of Fisher Scoring iterations: 2
```

Объем выборки завышен (180 наблюдений, вместо 18 субъектов). Стандартные ошибки и уровни значимости занижены. Увеличивается вероятность ошибок I рода.

Нарушено условие независимости наблюдений.

Плохое решение: не учитываем группирующий фактор

```
ggplot(sl, aes(x = Days, y = Reaction)) +
  geom_point() +
  geom_smooth(se = TRUE, method = "lm", size = 1)
```



Доверительная зона регрессии “заужена”.

Большие остатки, т.к. неучтенная межиндивидуальная изменчивость “ушла” в остаточную.

Громоздкое решение: группирующий фактор как фиксированный

$Reaction_i = \beta_0 + \beta_1 Days_i + \beta_2 Subject_{2\,i} + ... + \beta_2 Subject_{18\,i} + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma)$

Громоздкое решение: группирующий фактор как фиксированный

$Reaction_i = \beta_0 + \beta_1 Days_i + \beta_2 Subject_{2\,i} + ... + \beta_2 Subject_{18\,i} + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma)$

В матричном виде это можно записать так:

$$\begin{pmatrix} Reaction_1 \\ Reaction_2 \\ \vdots \\ Reaction_{180} \end{pmatrix} = \begin{pmatrix} 1 & Days_1 & Subject_{2\,1} & \cdots & Subject_{18\,1} \\ 1 & Days_2 & Subject_{2\,2} & \cdots & Subject_{18\,2} \\ \vdots & & & & \\ 1 & Days_{180} & Subject_{2\,180} & \cdots & Subject_{18\,180} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{180} \end{pmatrix}$$

То есть: **Reaction** $= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

Громоздкое решение: группирующий фактор как фиксированный

```r
W2 <- glm(Reaction ~ Days + Subject, data = sl)
coef(W2)
```

```
# (Intercept)        Days  Subject309  Subject310
#   295.03104    10.46729  -126.90085  -111.13256
#  Subject330  Subject331  Subject332  Subject333
#   -38.91241   -32.69778   -34.83176   -25.97552
#  Subject334  Subject335  Subject337  Subject349
#   -46.83178   -92.06379    33.58718   -66.29936
#  Subject350  Subject351  Subject352  Subject369
#   -28.53115   -52.03608    -4.71229   -36.09919
#  Subject370  Subject371  Subject372
#   -50.43206   -47.14979   -24.24770
```

Фрагмент `summary(W2)`:

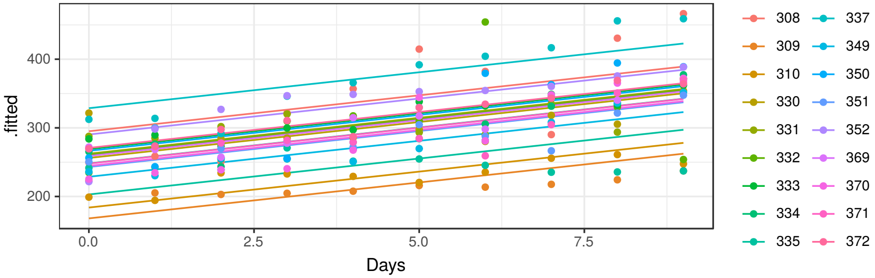
```
Residual standard error: 30.99 on 161 degrees of freedom
Multiple R-squared: 0.7277, Adjusted R-squared: 0.6973
F-statistic: 23.91 on 18 and 161 DF, p-value: < 2.2e-16
```

20 параметров (18 для `Subject`, один для `Days` и σ), а наблюдений всего 180.

Нужно минимум 10–20 наблюдений на каждый параметр (Harrell, 2013) — у нас всего 9.

Громоздкое решение: что нам делать с этим множеством прямых?

```
ggplot(fortify(W2), aes(x = Days, colour = Subject)) +
  geom_line(aes(y = .fitted, group = Subject)) +
  geom_point(data = sl, aes(y = Reaction)) +
  guides(colour = guide_legend(ncol = 2))
```



В модели, где субъект — фиксированный фактор, для каждого субъекта есть "поправка" для значения свободного члена в уравнении регрессии. Универсальность модели теряется: предсказания можно сделать только с учетом субъекта.

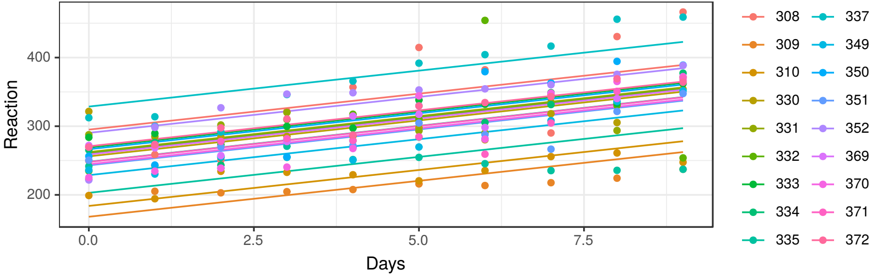
Фиксированные и случайные факторы

Фиксированные факторы

До сих пор мы имели дело только с фиксированными факторами.

Мы моделировали средние значения для уровней фиксированного фактора. Если групп было много, то приходилось моделировать много средних значений.

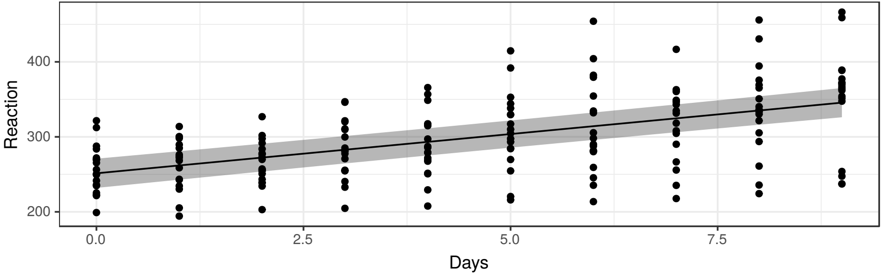
Поступая так, мы считали, что сравниваемые группы – фиксированные, и нам интересны именно сравнения между ними.



Можно посмотреть на группирующий фактор иначе!

Когда нам не важны конкретные значения интерсептов для разных уровней фактора, мы можем представить, что эффект фактора (величина “поправки”) — случайная величина, и можем оценить дисперсию между уровнями группирующего фактора.

Такие факторы называются **случайными факторами**.



Случайные факторы

- измерения в разные периоды времени
  - измерения, сделанные в химической лаборатории в разные дни
- измерения в разных участках пространства
  - урожай на участках одного поля
  - детали, произведенные на одном из нескольких аналогичных станков
- повторные измерения на одних и тех же субъектах
  - измерения до и после какого-то воздействия
- измерения на разных субъектах, которые сами объединены в группы
  - ученики в классах, классы в школах, школы в районах, районы в городах и т.п.

Случайные факторы в моделях

На один и тот же фактор можно посмотреть и как на фиксированный и как на случайный в зависимости от целей исследователя.
Поскольку моделируя случайный фактор мы оцениваем дисперсию между уровнями, то хорошо, если у случайного фактора будет минимум пять градаций.

GLMM со случайным отрезком

GLMM со случайным отрезком

$Reaction_{ij} = \beta_0 + \beta_1 Days_{ij} + c_i + \varepsilon_{ij}$

$c_i \sim N(0, \sigma_b)$ — случайный эффект субъекта (случайный отрезок, интерсепт)
$\varepsilon_{ij} \sim N(0, \sigma)$ — остатки модели
i — субъекты, j — дни

GLMM со случайным отрезком

$$Reaction_{ij} = \beta_0 + \beta_1 Days_{ij} + c_i + \varepsilon_{ij}$$

$c_i \sim N(0, \sigma_b)$ — случайный эффект субъекта (случайный отрезок, интерсепт)
$\varepsilon_{ij} \sim N(0, \sigma)$ — остатки модели
i — субъекты, j — дни

Для каждого субъекта i в матричном виде это записывается так:

$$\begin{pmatrix} Reaction_{i1} \\ Reaction_{i2} \\ \vdots \\ Reaction_{i10} \end{pmatrix} = \begin{pmatrix} 1 & Days_{i1} \\ 1 & Days_{i2} \\ \vdots & \\ 1 & Days_{i10} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} c_i + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{i10} \end{pmatrix}$$

То есть для всех субъектов: $\mathbf{Reaction} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon}$

Подберем модель со случайным отрезком

Используем `lmer` из пакета `lme4`.

```
?lmer # справка о lmer
```

`lmer` по умолчанию использует REML для подбора параметров. Это значит, что случайные эффекты будут оценены более точно, чем при использовании ML.

```
M1 <- lmer(Reaction ~ Days + (1 | Subject), data = sl)
```

Запишем уравнение модели со случайным отрезком

```r
summary(M1)
```

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Reaction ~ Days + (1 | Subject)
#    Data: sl
#
# REML criterion at convergence: 1786.5
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -3.2257 -0.5529  0.0109  0.5188  4.2506
#
# Random effects:
#  Groups   Name        Variance Std.Dev.
#  Subject  (Intercept) 1378.2   37.12
#  Residual              960.5   30.99
# Number of obs: 180, groups:  Subject, 18
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept) 251.4051     9.7467   25.79
# Days         10.4673     0.8042   13.02
#
# Correlation of Fixed Effects:
#      (Intr)
# Days -0.371
```

$$Reaction_{ij} = 251.4 + 10.5 Days_{ij} + c_i + \varepsilon_{ij}$$

$c_i \sim N(0, 37.12)$ — случайный эффект субъекта
$\varepsilon_{ij} \sim N(0, 30.99)$ — остатки модели
i — субъекты, j — дни

Предсказания смешанных моделей бывают двух типов

- Предсказания с учетом лишь фиксированных эффектов,
- Предсказания с одновременным учетом как фиксированных, так и случайных эффектов.

Данные для графика предсказаний фиксированной части модели:

```
library(dplyr)
NewData <- sl %>% group_by(Subject) %>%
  do(data.frame(Days = seq(min(.$Days), max(.$Days), length = 10)))

head(NewData, 3)
```

```
# # A tibble: 3 x 2
# # Groups:   Subject [1]
#   Subject  Days
#   <fct>   <dbl>
# 1 308         0
# 2 308         1
# 3 308         2
```

Предсказания фиксированной части модели при помощи predict()

```
?predict.merMod
```

Функция `predict()` в `lme4` не считает стандартные ошибки и доверительные интервалы. Это потому, что нет способа адекватно учесть неопределенность, связанную со случайными эффектами.

```
NewData$fit <- predict(M1, NewData, type = 'response', re.form = NA)
head(NewData, 3)
```

```
# # A tibble: 3 x 3
# # Groups:   Subject [1]
#   Subject  Days   fit
#   <fct>   <dbl> <dbl>
# 1 308         0  251.
# 2 308         1  262.
# 3 308         2  272.
```

Предсказания фиксированной части модели в матричном виде

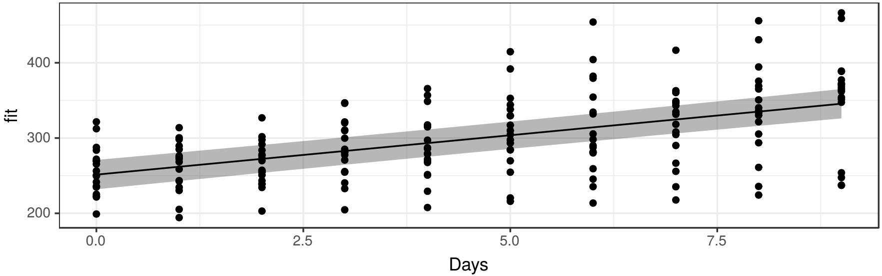
Стандартные ошибки, рассчитанные обычным методом,
позволят получить **приблизительные** доверительные интервалы.

```
# Предсказанные значения при помощи матриц
X <- model.matrix(~ Days, data = NewData)
betas <- fixef(M1)
NewData$fit <- X %*% betas

# Стандартные ошибки
NewData$SE <- sqrt( diag(X %*% vcov(M1) %*% t(X)) )

NewData$lwr <- NewData$fit - 2 * NewData$SE
NewData$upr <- NewData$fit + 2 * NewData$SE
```

Более точные доверительные интервалы можно получить при помощи бутстрепа.
Мы сделаем это позже для финальной модели.

График предсказаний фиксированной части модели

```
ggplot(data = NewData, aes(x = Days, y = fit)) +
  geom_ribbon(alpha = 0.35, aes(ymin = lwr, ymax = upr)) +
  geom_line() + geom_point(data = sl, aes(x = Days, y = Reaction))
```

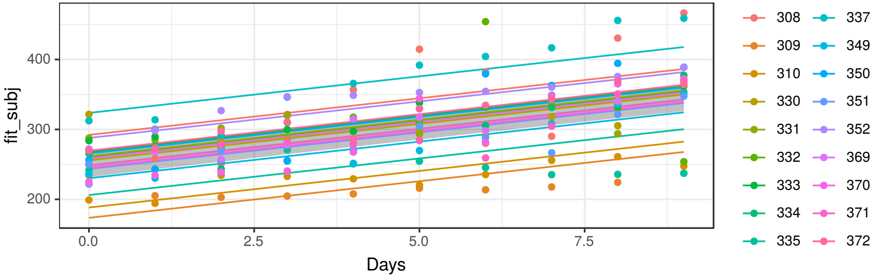


Зависимость времени реакции от продолжительности периода бессонницы без учета субъекта:

$$\widehat{Reaction}_{ij} = 251.4 + 10.5 Days_{ij}$$

Предсказания для каждого уровня случайного фактора

```
NewData$fit_subj <- predict(M1, NewData, type = 'response')
ggplot(NewData, aes(x = Days, y = fit_subj)) +
  geom_ribbon(alpha = 0.3, aes(ymin = lwr, ymax = upr)) +
  geom_line(aes(colour = Subject)) +
  geom_point(data = sl, aes(x = Days, y = Reaction, colour = Subject))  +
  guides(colour = guide_legend(ncol = 2))
```



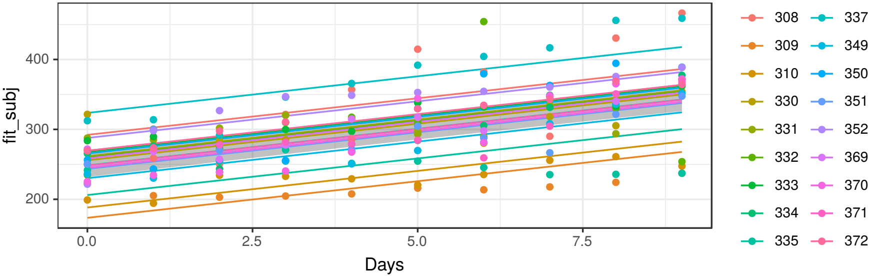
Зависимость времени реакции от продолжительности периода бессонницы для обследованных субъектов:

$$\widehat{Reaction}_{ij} = 251.4 + 10.5 Days_{ij} + c_i$$

Важные замечания

Случайный фактор помогает учесть взаимозависимость наблюдений для каждого из субъектов – “индуцированные” корреляции.

После анализа остатков модели можно будет понять, стоит ли с ней работать дальше. Одна из потенциальных проблем – время реакции разных субъектов может меняться непараллельно. Возможно, модель придется переформулировать.



Индуцированная корреляция

Разберемся со случайной частью модели

Reaction $= \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon}$
b $\sim N(0, \mathbf{D})$ - случайные эффекты b_i нормально распределены со средним 0 и матрицей ковариаций **D**
$\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$ - остатки модели нормально распределены со средним 0 и матрицей ковариаций $\boldsymbol{\Sigma}$

Разберемся со случайной частью модели

Reaction $= \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon}$
b $\sim N(0, \mathbf{D})$ - случайные эффекты b_i нормально распределены со средним 0 и матрицей ковариаций **D**
$\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$ - остатки модели нормально распределены со средним 0 и матрицей ковариаций $\boldsymbol{\Sigma}$

Матрица ковариаций остатков: $\boldsymbol{\Sigma} = \sigma^2 \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$

Остатки модели должны быть независимы друг от друга

В матрице ковариаций остатков вне диагонали стоят нули, т.е. ковариация разных остатков равна нулю. Т.е. остатки независимы друг от друга.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

В то же время, отдельные значения переменной-отклика **Y** уже не будут независимы друг от друга при появлении в модели случайного фактора.

Матрица ковариаций переменной-отклика

$\textbf{Reaction} = \mathbf{X}\beta + \mathbf{Zb} + \varepsilon$
$\mathbf{b} \sim N(0, \mathbf{D})$
$\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$

Можно показать, что переменная-отклик **Y** нормально распределена:

$\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{V})$

Матрица ковариаций переменной-отклика

$\textbf{Reaction} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon}$
$\mathbf{b} \sim N(0, \mathbf{D})$
$\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$

Можно показать, что переменная-отклик **Y** нормально распределена:

$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$

Матрица ковариаций переменной-отклика:

$\mathbf{V} = \mathbf{ZDZ}' + \boldsymbol{\Sigma}$

где **D** — матрица ковариаций случайных эффектов.

Т.е. **добавление случайных эффектов приводит к изменению ковариационной матрицы V**

Добавление случайных эффектов приводит к изменению ковариационной матрицы

$$\mathbf{V} = \mathbf{ZDZ}' + \mathbf{\Sigma}$$

Для простейшей смешанной модели со случайным отрезком:

$$\mathbf{V} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \cdot \sigma_b^2 \cdot \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} + \sigma^2 \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 + \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma^2 + \sigma_b^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma^2 + \sigma_b^2 \end{pmatrix}$$

Индуцированная корреляция — следствие включения в модель случайных эффектов

$$\mathbf{V} = \begin{pmatrix} \sigma^2 + \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma^2 + \sigma_b^2 & \cdots & \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma^2 + \sigma_b^2 \end{pmatrix}$$

- σ_b^2 — ковариация между наблюдениями одного субъекта.
- $\sigma^2 + \sigma_b^2$ — дисперсия.

Т.е. корреляция между наблюдениями одного субъекта $\sigma_b^2/(\sigma^2 + \sigma_b^2)$

Коэффициент внутриклассовой корреляции (intra-class correlation, ICC)

$ICC = \sigma_b^2/(\sigma^2 + \sigma_b^2)$

Способ измерить, насколько коррелируют друг с другом наблюдения из одной и той же группы, заданной случайным фактором.

Если ICC низок, то наблюдения очень разные внутри каждой из групп. Значит, чтобы надежно оценить эффект этого случайного фактора, нужно брать больше наблюдений в группе.

Если ICC высок, то наблюдения очень похожи внутри каждой из групп, заданных случайным фактором. Значит, можно брать меньше наблюдений в группе.

Вычислим коэффициент внутриклассовой корреляции

```r
summary(M1)
```

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Reaction ~ Days + (1 | Subject)
#    Data: sl
#
# REML criterion at convergence: 1786.5
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -3.2257 -0.5529  0.0109  0.5188  4.2506
#
# Random effects:
#  Groups   Name        Variance Std.Dev.
#  Subject  (Intercept) 1378.2   37.12
#  Residual              960.5   30.99
# Number of obs: 180, groups:  Subject, 18
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept) 251.4051     9.7467   25.79
# Days         10.4673     0.8042   13.02
#
# Correlation of Fixed Effects:
#      (Intr)
# Days -0.371
```

```r
VarCorr(M1)  # Случайные эффекты отдельно
```

```
#  Groups   Name        Std.Dev.
#  Subject  (Intercept) 37.124
#  Residual             30.991
```

```r
37.124^2 / (37.124^2 + 30.991^2)
```

```
# [1] 0.5893148
```

Диагностика модели

Условия применимости

- Случайность и независимость наблюдений.
- Линейная связь.
- Нормальное распределение остатков.
- Гомогенность дисперсий остатков.
- Отсутствие коллинеарности предикторов.

Данные для анализа остатков

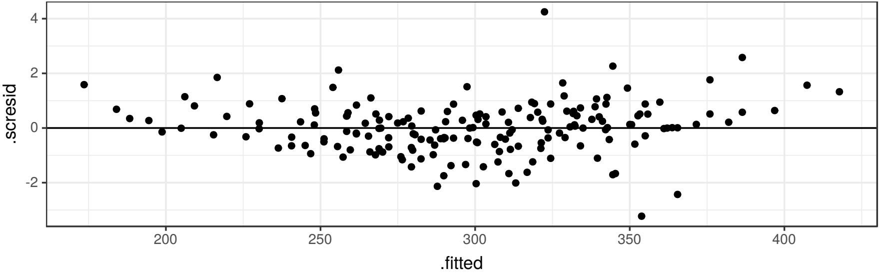
```
M1_diag <- data.frame(
  sl,
  .fitted = predict(M1),
  .resid = resid(M1, type = 'pearson'),
  .scresid = resid(M1, type = 'pearson', scaled = TRUE))

head(M1_diag, 4)
```

```
#   Reaction Days Subject  .fitted     .resid    .scresid
# 1 249.5600    0     308 292.1888 -42.628815 -1.37551202
# 2 258.7047    1     308 302.6561 -43.951401 -1.41818815
# 3 250.8006    2     308 313.1234 -62.322787 -2.01098113
# 4 321.4398    3     308 323.5907  -2.150873 -0.06940261
```

- .fitted — предсказанные значения.
- .resid — Пирсоновские остатки.
- .scresid — стандартизованные Пирсоновские остатки.

График остатков от предсказанных значений

```
gg_resid <- ggplot(M1_diag, aes(y = .scresid)) +
  geom_hline(yintercept = 0)
gg_resid + geom_point(aes(x = .fitted))
```

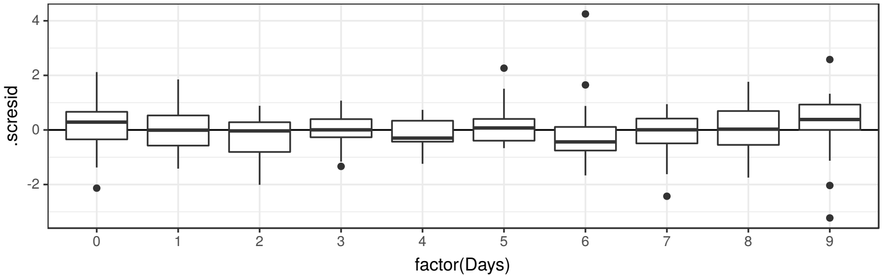


Большие остатки.

Гетерогенность дисперсий.

Графики остатков от ковариат в модели и не в модели

```
gg_resid + geom_boxplot(aes(x = factor(Days)))
```

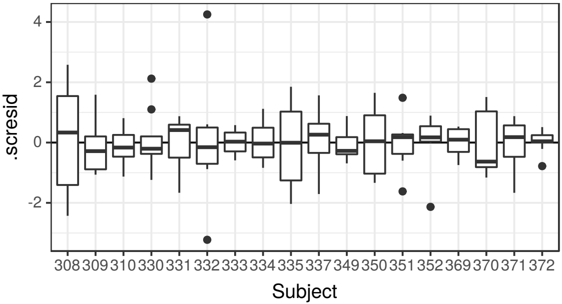


Большие остатки в некоторые дни.

Гетерогенность дисперсий.

Графики остатков от ковариат в модели и не в модели

```
gg_resid + geom_boxplot(aes(x = Subject))
```



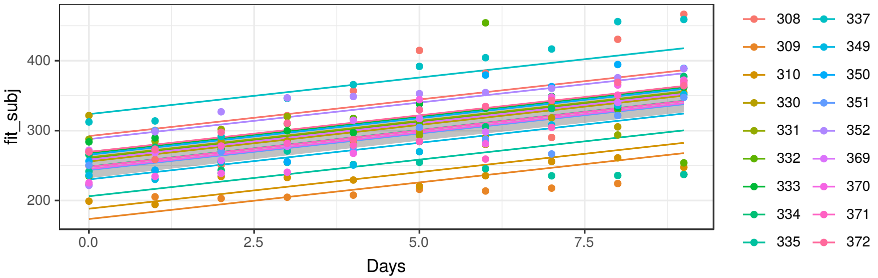
Большие остатки у 332 субъекта.

Гетерогенность дисперсий.

GLMM со случайным отрезком и углом наклона

GLMM со случайным отрезком и углом наклона

На графике индивидуальных эффектов было видно, что измерения для разных субъектов, возможно, идут непараллельно. Усложним модель — добавим случайные изменения угла наклона для каждого из субъектов.



Это можно биологически объяснить. Возможно, в зависимости от продолжительности бессонницы у разных субъектов скорость реакции будет ухудшаться разной скоростью: одни способны выдержать 9 дней почти без потерь, а другим уже пары дней может быть достаточно.

Уравнение модели со случайным отрезком и углом наклона

$$Reaction_{ij} = \beta_0 + \beta_1 Days_{ij} + c_i + d_{ij} Days_{ij} + \varepsilon_{ij}$$

$c_i \sim N(0, \sigma_b)$ — случайный интерсепт для субъекта
$d_{ij} \sim N(0, \sigma_c)$ — случайный угол наклона для субъекта
$\varepsilon_{ij} \sim N(0, \sigma)$ — остатки модели
i — субъекты, j — дни

Уравнение модели со случайным отрезком и углом наклона

$Reaction_{ij} = \beta_0 + \beta_1 Days_{ij} + c_i + d_{ij} Days_{ij} + \varepsilon_{ij}$

$c_i \sim N(0, \sigma_b)$ — случайный интерсепт для субъекта
$d_{ij} \sim N(0, \sigma_c)$ — случайный угол наклона для субъекта
$\varepsilon_{ij} \sim N(0, \sigma)$ — остатки модели
i — субъекты, j — дни

Для каждого субъекта i в матричном виде это записывается так:

$$\begin{pmatrix} Reaction_{i1} \\ Reaction_{i2} \\ \vdots \\ Reaction_{i10} \end{pmatrix} = \begin{pmatrix} 1 & Days_{i1} \\ 1 & Days_{i2} \\ \vdots & \\ 1 & Days_{i10} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 & Days_{i1} \\ 1 & Days_{i2} \\ \vdots & \\ 1 & Days_{i10} \end{pmatrix} \begin{pmatrix} c_i \\ d_i \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{i10} \end{pmatrix}$$

То есть для всех субъектов: $\mathbf{Reaction} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\varepsilon}$

Подберем модель со случайным отрезком и углом наклона

Формат записи формулы для случайных эффектов в `lme4`

```
(1 + угловой_коэффициент | отрезок)
```

```
MS1 <- lmer(Reaction ~ Days + ( 1 + Days|Subject), data = sl)
```

Запишем уравнение модели со случайным отрезком

```r
summary(MS1)
```

```
# Linear mixed model fit by REML ['lmerMod']
# Formula: Reaction ~ Days + (1 + Days | Subject)
#    Data: sl
#
# REML criterion at convergence: 1743.6
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -3.9536 -0.4634  0.0231  0.4634  5.1793
#
# Random effects:
#  Groups   Name        Variance Std.Dev. Corr
#  Subject  (Intercept) 612.09   24.740
#           Days         35.07    5.922   0.07
#  Residual             654.94   25.592
# Number of obs: 180, groups:  Subject, 18
#
# Fixed effects:
#             Estimate Std. Error t value
# (Intercept)  251.405      6.825  36.838
# Days          10.467      1.546   6.771
#
# Correlation of Fixed Effects:
#      (Intr)
# Days -0.138
```

$Reaction_{ij} = 251.4 + 10.5 Days_{ij} + c_i + d_{ij} Days_{ij} + \varepsilon_{ij}$

$c_i \sim N(0, 24.74)$ — случайный интерсепт для субъекта

$d_{ij} \sim N(0, 5.92)$ — случайный угол наклона для субъекта

$\varepsilon_{ij} \sim N(0, 25.59)$ — остатки модели

i — субъекты, j — дни

Данные для графика предсказаний фиксированной части модели

```
library(dplyr)
NewData <- sl %>% group_by(Subject) %>%
  do(data.frame(Days = seq(min(.$Days), max(.$Days), length = 10)))

NewData$fit <- predict(MS1, NewData, type = 'response', re.form = NA)
head(NewData, 3)
```

```
# # A tibble: 3 x 3
# # Groups:   Subject [1]
#   Subject  Days   fit
#   <fct>   <dbl> <dbl>
# 1 308         0  251.
# 2 308         1  262.
# 3 308         2  272.
```

Предсказания фиксированной части модели в матричном виде

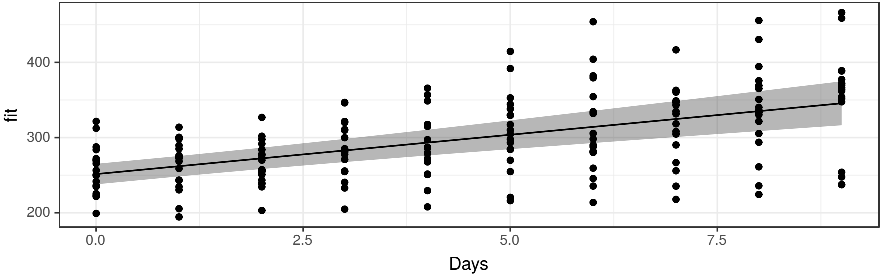
Вычислим **приблизительные** доверительные интервалы.

```
# Предсказанные значения при помощи матриц
X <- model.matrix(~ Days, data = NewData)
betas <- fixef(MS1)
NewData$fit <- X %*% betas

# Стандартные ошибки
NewData$SE <- sqrt( diag(X %*% vcov(MS1) %*% t(X)) )

NewData$lwr <- NewData$fit - 2 * NewData$SE
NewData$upr <- NewData$fit + 2 * NewData$SE
```

Более точные доверительные интервалы можно получить при помощи бутстрепа.

График предсказаний фиксированной части модели

```r
ggplot(data = NewData, aes(x = Days, y = fit)) +
  geom_ribbon(alpha = 0.35, aes(ymin = lwr, ymax = upr)) +
  geom_line() + geom_point(data = sl, aes(x = Days, y = Reaction))
```

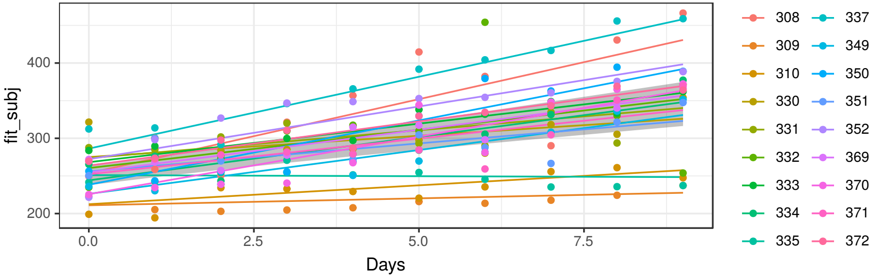


Зависимость времени реакции от продолжительности периода бессонницы без учета субъекта:

$$\widehat{Reaction}_{ij} = 251.4 + 10.5 Days_{ij}$$

Предсказания для каждого уровня случайного фактора

```
NewData$fit_subj <- predict(MS1, NewData, type = 'response')
ggplot(NewData, aes(x = Days, y = fit_subj)) +
  geom_ribbon(alpha = 0.3, aes(ymin = lwr, ymax = upr)) +
  geom_line(aes(colour = Subject)) +
  geom_point(data = sl, aes(x = Days, y = Reaction, colour = Subject))  +
  guides(colour = guide_legend(ncol = 2))
```



Зависимость времени реакции от продолжительности периода бессонницы для обследованных субъектов:

$$\widehat{Reaction}_{ij} = 251.4 + 10.5 Days_{ij} + c_i + d_{ij} Days_{ij}$$

Диагностика модели со случайным отрезком и углом наклона

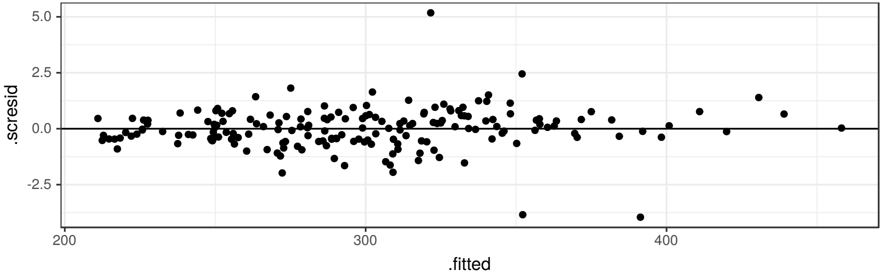
Данные для анализа остатков

```r
MS1_diag <- data.frame(
  sl,
  .fitted = predict(MS1),
  .resid = resid(MS1, type = 'pearson'),
  .scresid = resid(MS1, type = 'pearson', scaled = TRUE))

head(MS1_diag, 4)
```

```
#   Reaction Days Subject  .fitted     .resid   .scresid
# 1 249.5600    0     308 253.6637  -4.103670 -0.1603509
# 2 258.7047    1     308 273.3299 -14.625228 -0.5714807
# 3 250.8006    2     308 292.9962 -42.195586 -1.6487922
# 4 321.4398    3     308 312.6624   8.777356  0.3429751
```

График остатков от предсказанных значений

```
gg_resid <- ggplot(MS1_diag, aes(y = .scresid)) +
  geom_hline(yintercept = 0)
gg_resid + geom_point(aes(x = .fitted))
```

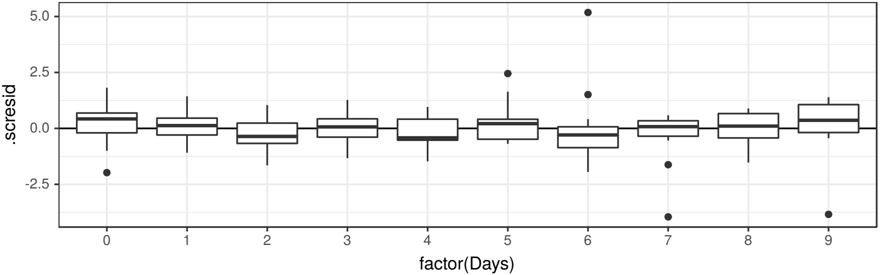


Несколько больших остатков.

Гетерогенность дисперсий не выражена.

Графики остатков от ковариат в модели и не в модели

```
gg_resid + geom_boxplot(aes(x = factor(Days)))
```

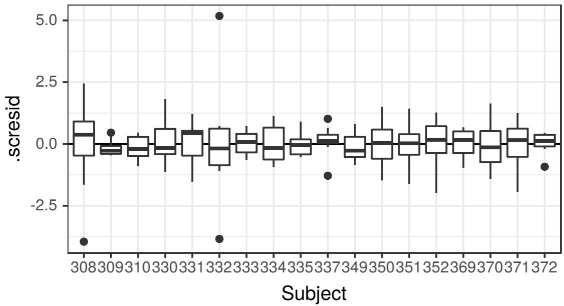


Большие остатки в некоторые дни.

Нет гетерогенности дисперсий остатков.

Графики остатков от ковариат в модели и не в модели

```
gg_resid + geom_boxplot(aes(x = Subject))
```



Большие остатки у 332 субъекта.

Гетерогенность дисперсий не выражена.

Смешанные линейные модели

Смешанные модели (Mixed Models)

Смешанными называются модели, включающие случайные факторы.

- Общие смешанные модели (General Linear Mixed Models) — только нормальное распределение зависимой переменной.
- Обобщенные смешанные модели (Generalized Linear Mixed Models) — распределения зависимой переменной могут быть другими (из семейства экспоненциальных распределений).

Смешанная линейная модель в общем виде

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Zb} + \varepsilon$$

$\mathbf{b} \sim N(0, \mathbf{D})$ — случайные эффекты нормально распределены со средним 0 и матрицей ковариаций $\mathbf{D}$ (ее диагональные элементы – стандартное отклонение σ_b).

$\boldsymbol{\varepsilon} \sim N(0, \mathbf{\Sigma})$ — остатки модели нормально распределены со средним 0 и матрицей ковариаций $\mathbf{\Sigma}$ (ее диагональные элементы – стандартное отклонение σ).

$\mathbf{X}\beta$ — фиксированная часть модели.

$\mathbf{Zb}$ — случайная часть модели.

В зависимости от устройства модельной матрицы для случайных эффектов **Z** смешанные модели делят на модели со случайным отрезком и случайным углом наклона.

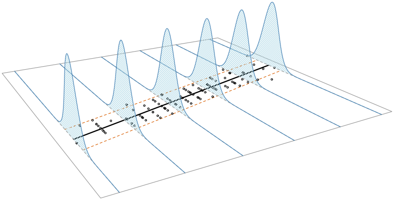
Методы подбора параметров в смешанных моделях

Метод максимального правдоподобия (Maximum Likelihood, ML)

Метод ограниченного максимального правдоподобия (Restricted Maximum Likelihood, REML)

Метод максимального правдоподобия, ML

Правдоподобие (likelihood) — способ измерить соответствие имеющихся данных тому, что можно получить при определенных значениях параметров модели.



Это произведение вероятностей получения каждой из точек данных:

$$L(\theta|\text{data}) = \Pi_{i=1}^{n} f(\text{data}|\theta)$$

- $f(\text{data}|\theta)$ — функция распределения с параметрами θ

Параметры модели должны максимизировать значение логарифма правдоподобия

$$lnL(\theta|\text{data}) \longrightarrow \text{max}$$

ML-оценки для дисперсий – смещенные

Например, ML оценка дисперсии будет смещенной:

$$\hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n},$$

т.к. в знаменателе не $n - 1$, а n.

Аналогичные проблемы возникают при ML оценках для случайных эффектов в линейных моделях.

Это происходит потому, что в смешанной линейной модели

$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}$

$Y \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, где $\mathbf{ZDZ}' + \mathbf{\Sigma}$

т.е. одновременно приходится оценивать β и $\mathbf{V}$.

Метод ограниченного максимального правдоподобия, REML

$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Zb} + \varepsilon$

$Y \sim N(\mathbf{X}\beta, \mathbf{V})$, где $\mathbf{ZDZ}' + \mathbf{\Sigma}$

т.е. одновременно приходится оценивать β и **V**.

Если найти матрицу **A**, ортогональную к **X** (т.е. $\mathbf{AX} = 0$), то умножив ее на Y можно избавиться от $\boldsymbol{\beta}$:

$\mathbf{A}'\mathbf{Y} = \mathbf{A}'\mathbf{X}\beta + \mathbf{A}'\mathbf{V} = \mathbf{0} + \mathbf{A}'\mathbf{V}$

$A'Y \sim N(\mathbf{0}, \mathbf{A}'\mathbf{VA})$

Тогда можно воспользоваться ML, чтобы найти V.

В результате получатся несмещенные оценки дисперсий, но немного другие оценки β.

ML или REML

Если нужно работать с правдоподобиями – ML

Если нужны точные оценки случайных эффектов – REML

Тестирование гипотез в смешанных моделях

Использование смешанных моделей для получения выводов

Тесты, которые традиционно применяются для GLM, дадут лишь **приблизительные результаты** для GLMM:

- t-(или z-) тесты Вальда для коэффициентов,
- тесты отношения правдоподобий (Likelihood ratio tests, LRT).

Поэтому для отбора моделей применяют подход, не связанный с тестами:

- информационные критерии (AIC, BIC и т.п.).

Наиболее точные результаты тестов можно получить, используя **"золотой стандарт"**:

- параметрический бутстреп.

t-(или -z) тесты Вальда

$H_0 : \beta_k = 0$, $\quad H_A : \beta_k \neq 0$

$$\frac{b_k}{SE_{b_k}} \sim N(0, 1) \quad \text{или} \quad \frac{b_k}{SE_{b_k}} \sim t_{(df=n-p)}$$, если нужно оценивать σ

b_k — оценка коэффициента, n — объем выборки, p — число параметров модели.

t-(или -z) тесты Вальда дают лишь приблизительный результат,
поэтому в пакете `lme4` даже не приводят уровни значимости в `summary()`.
Не рекомендуется ими пользоваться.

```
coef(summary(MS1))
```

```
#               Estimate Std. Error   t value
# (Intercept) 251.40510   6.824556 36.838310
# Days         10.46729   1.545789  6.771485
```

Тесты отношения правдоподобий (LRT)

$$LRT = 2ln\Big(\frac{L_{M_1}}{L_{M_2}}\Big) = 2(lnL_{M_1} - lnL_{M_2})$$

- M_1 и M_2 — вложенные модели (M_1 — более полная, M_2 — уменьшенная),
- L_{M_1}, L_{M_2} – правдоподобия моделей и lnL_{M_1}, lnL_{M_2} – логарифмы правдоподобий.

Распределение LRT **аппроксимируют** распределением χ^2 с $df = df_{M_2} - df_{M_1}$.

- LRT консервативен для случайных эффектов, т.к. тест гипотезы вида $H_0 : \hat{\sigma}_k^2 = 0$ происходит на границе области возможных значений параметра.
- LRT либерален для фиксированных эффектов, дает заниженные уровни значимости.

LRT для случайных эффектов

Модели **с одинаковой фиксированной частью**, подобранные REML. Уровни значимости будут завышены.

```
MS1 <- lmer(Reaction ~ Days + (1 + Days | Subject), data = sl, REML = TRUE)
MS0 <- lmer(Reaction ~ Days + (1 | Subject), data = sl, REML = TRUE)
anova(MS1, MS0, refit = FALSE)
```

```
# Data: sl
# Models:
# MS0: Reaction ~ Days + (1 | Subject)
# MS1: Reaction ~ Days + (1 + Days | Subject)
#     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
# MS0  4 1794.5 1807.2 -893.23   1786.5
# MS1  6 1755.6 1774.8 -871.81   1743.6 42.837      2   4.99e-10 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Время реакции у разных людей по-разному зависит от продолжительности бессонницы.

Обычно тесты не делают, набор случайных эффектов определяется устройством данных.

LRT для фиксированных эффектов

Модели **с одинаковой случайной частью**, подобранные ML. Уровни значимости будут занижены.

```
MS1.ml <- lmer(Reaction ~ Days + (1 + Days | Subject), data = sl, REML = FALSE)
MS0.ml <- lmer(Reaction ~ 1 + (1 + Days | Subject), data = sl, REML = FALSE)
anova(MS1.ml, MS0.ml)
```

```
# Data: sl
# Models:
# MS0.ml: Reaction ~ 1 + (1 + Days | Subject)
# MS1.ml: Reaction ~ Days + (1 + Days | Subject)
#        Df    AIC    BIC  logLik deviance  Chisq Chi Df  Pr(>Chisq)
# MS0.ml  5 1785.5 1801.4 -887.74   1775.5
# MS1.ml  6 1763.9 1783.1 -875.97   1751.9 23.537      1 0.000001226
#
# MS0.ml
# MS1.ml ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Время реакции зависит от продолжительности бессонницы.

Сравнение моделей по AIC

Можно сравнить вложенные или невложенные модели, подобранные ML, с одинаковой случайной частью.

```
AIC(MS1.ml, MS0.ml)
```

```
#         df      AIC
# MS1.ml  6 1763.939
# MS0.ml  5 1785.476
```

Время реакции зависит от продолжительности бессонницы (AIC).

Бутстреп для тестирования значимости и для предсказаний

Параметрический бутстреп для LRT

Чтобы при помощи бутстрепа получить оценку уровня значимости для LRT при сравнении двух моделей M_{full} и $M_{reduced}$, нужно

1.Многократно повторить:
- сгенерировать новые данные из уменьшенной модели, - по сгенерированным данным подобрать полную и уменьшенную модели и рассчитать LRT.

2.Построить распределение LRT по всем итерациям бутстрепа.

Уровень значимости – это доля итераций, в которых получено LRT больше, чем данное.

Параметрический бутстреп для LRT фиксированных эффектов

В строке PBtest – значение LRT и его уровень значимости, полученный бутстрепом.

```
library(pbkrtest)
pmod <- PBmodcomp(MS1.ml, MS0.ml, nsim = 100) # 1000 и больше для реальных данных
summary(pmod)
```

```
# Parametric bootstrap test; time: 4.91 sec; samples: 100 extremes: 0;
# large : Reaction ~ Days + (1 + Days | Subject)
# small : Reaction ~ 1 + (1 + Days | Subject)
#             stat     df    ddf      p.value
# PBtest   23.537                   0.0099010 **
# Gamma    23.537                0.0000001092 ***
# Bartlett 20.502  1.000         0.0000059575 ***
# F        23.537  1.000 15.511    0.0001926 ***
# LRT      23.537  1.000         0.0000012256 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Бутстреп-оценка доверительной зоны регрессии

Чтобы при помощи бутстрепа оценить положение зоны, где с 95% вероятностью будут лежать предсказанные значения, нужно:

1.Многократно повторить:
- сгенерировать новые данные из модели - по сгенерированным данным подобрать модель и получить ее предсказания

2.Построить распределение предсказанных значений по всем итерациям бутстрепа.

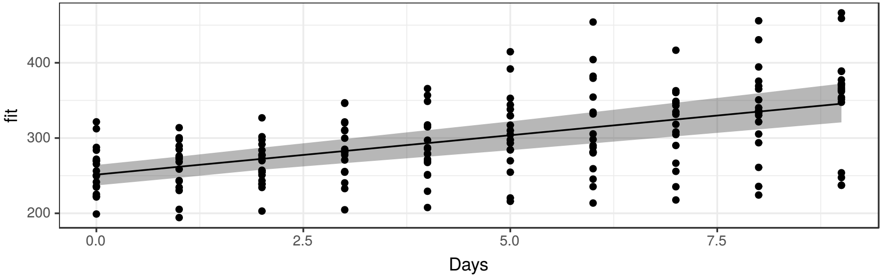
95% доверительная зона регрессии — это область, куда попали предсказанные значения в 95% итераций (т.е. между 0.025 и 0.975 персентилями).

Бутстреп-оценка доверительной зоны регрессии

```
NewData <- sl %>% group_by(Subject) %>%
  do(data.frame(Days = seq(min(.$Days), max(.$Days), length = 10)))
NewData$fit <- predict(MS1, NewData, type = 'response', re.form = NA)

# Многократно симулируем данные из модели и получаем для них предсказанные значения
bMS1 <- bootMer(x = MS1,
                FUN = function(x) predict(x, new_data = NewData, re.form = NA),
                nsim = 100)

# Рассчитываем квантили предсказанных значений для всех итераций бутстрепа
b_se <- apply(X = bMS1$t,
              MARGIN = 2,
              FUN = function(x) quantile(x, probs = c(0.025, 0.975), na.rm = TRUE))

# Доверительная зона для предсказанных значений
NewData$lwr <- b_se[1, ]
NewData$upr <- b_se[2, ]
```

График предсказаний фиксированной части модели

```
ggplot(data = NewData, aes(x = Days, y = fit)) +
  geom_ribbon(alpha = 0.35, aes(ymin = lwr, ymax = upr)) +
  geom_line() + geom_point(data = sl, aes(x = Days, y = Reaction))
```



Take-home messages

Свойства	Фиксированные факторы	Случайные факторы
Уровни фактора	фиксированные, заранее определенные и потенциально воспроизводимые уровни	случайная выборка из всех возможных уровней
Используются для тестирования гипотез	о средних значениях отклика между уровнями фактора $H_0 : \mu_1 = \mu_2 = ... = \mu_i = \mu$	о дисперсии отклика между уровнями фактора $H_0 : \sigma^2_{rand.fact.} = 0$
Выводы можно экстраполировать	только на уровни из анализа	на все возможные уровни
Число уровней фактора	Осторожно! Если уровней фактора слишком много, то нужно подбирать слишком много коэффициентов — должно быть много данных	Важно! Для точной оценки σ нужно нужно много уровней фактора — не менее 5

Take-home messages

- Смешанные модели могут включать случайные и фиксированные факторы.
    - Градации фиксированных факторов заранее определены, а выводы можно экстраполировать только на такие уровни, которые были задействованы в анализе. Тестируется гипотеза о равенстве средних в группах.
    - Градации случайных факторов — выборка из возможных уровней, а выводы можно экстраполировать на другие уровни. Тестируется гипотеза о дисперсии между группами.
- Есть два способа подбора коэффициентов в смешанных моделях: ML и REML. Для разных этапов анализа важно, каким именно способом подобрана модель.
- Коэффициент внутриклассовой корреляции оценивает, насколько коррелируют друг с другом наблюдения из одной и той же группы случайного фактора.
- Случайные факторы могут описывать вариацию как интерсептов, так и коэффициентов угла наклона.
- Модели со смешанными эффектами позволяют получить предсказания как общем уровне, так и на уровне отдельных субъектов.

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