

Смешанные линейные модели

Линейные модели...

Марина Варфоломеева, Вадим Хайтов

СПбГУ



Вы узнаете

- ▶ Что такое смешанные модели и когда они применяются
- ▶ Что такое фиксированные и случайные факторы

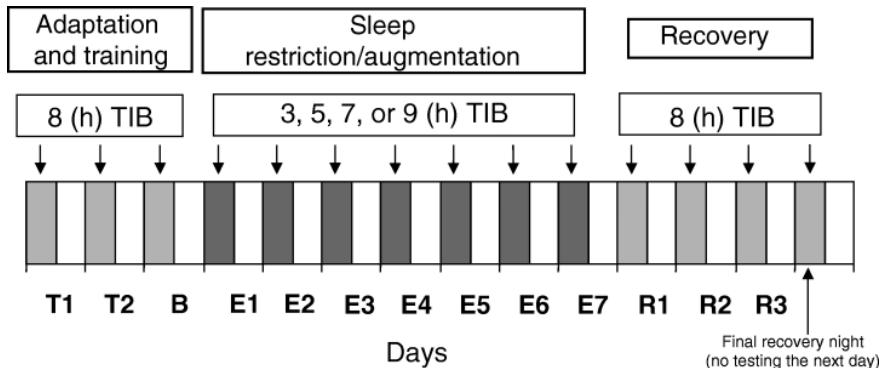
Вы сможете

- ▶ 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 смешанную линейную модель со случайным отрезком и случайным углом наклона в R при помощи методов максимального правдоподобия

“Многоуровневые” данные

Пример: Как время реакции людей зависит от бессонницы?

В статье Belenky et al., 2003. приводится такая схема исследования:



В датасете `sleepstudy` из пакета `lme4` описание немного отличается от того, что в статье: В ночь перед нулевым днем всем испытуемым давали поспать нормальное время, а в следующие 9 ночей — давали спать по 3 часа. Каждый день измеряли время реакции в серии тестов.

Данные: Belenky et al. (2003) Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. *Journal of Sleep Research* 12, 1–12.



Данные sleepstudy

- ▶ 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
head(sl, 3)
```

```
#   Reaction Days Subject
# 1 249.5600    0     308
# 2 258.7047    1     308
# 3 250.8006    2     308
```

Знакомство с данными

```
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
```

```
# пропущенные значения
```

```
colSums(is.na(sl))
```

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

Знакомство с данными

```
# число субъектов
```

```
length(unique(sl$Subject))
```

```
# [1] 18
```

```
# сбалансирован ли объем выборки?
```

```
table(sl$Subject)
```

```
#
```

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

```
# 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
```

```
table(sl$Subject, sl$Days)
```

```
#
```

```
#      0 1 2 3 4 5 6 7 8 9
```

```
# 308 1 1 1 1 1 1 1 1 1 1
```

```
# 309 1 1 1 1 1 1 1 1 1 1
```

```
# 310 1 1 1 1 1 1 1 1 1 1
```

```
# 330 1 1 1 1 1 1 1 1 1 1
```

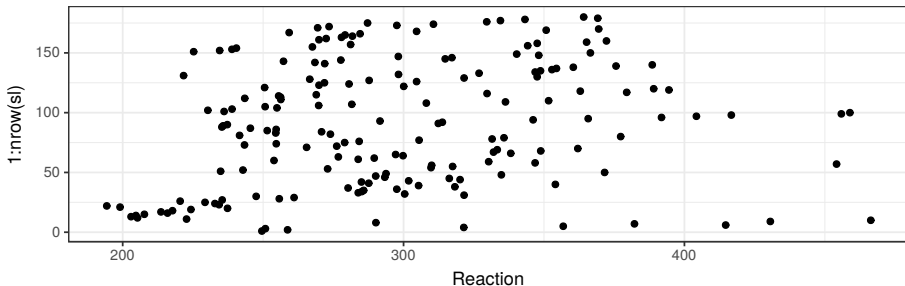
```
# 331 1 1 1 1 1 1 1 1 1 1
```

```
# 332 1 1 1 1 1 1 1 1 1 1
```

```
# 333 1 1 1 1 1 1 1 1 1 1
```

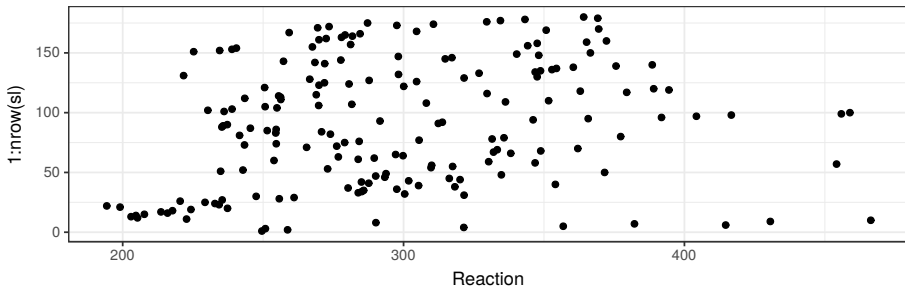
Есть ли выбросы?

```
library(ggplot2)
theme_set(theme_bw())
# построим дот-плот
ggplot(sl, aes(x = Reaction, y = 1:nrow(sl))) +
  geom_point()
```



Есть ли выбросы?

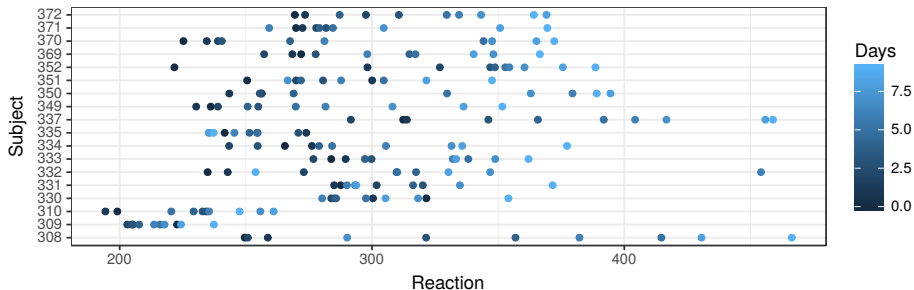
```
library(ggplot2)
theme_set(theme_bw())
# построим дот-плот
ggplot(sl, aes(x = Reaction, y = 1:nrow(sl))) +
  geom_point()
```



Кажется, что нет ничего странного, но мы еще не учли информацию о субъектах

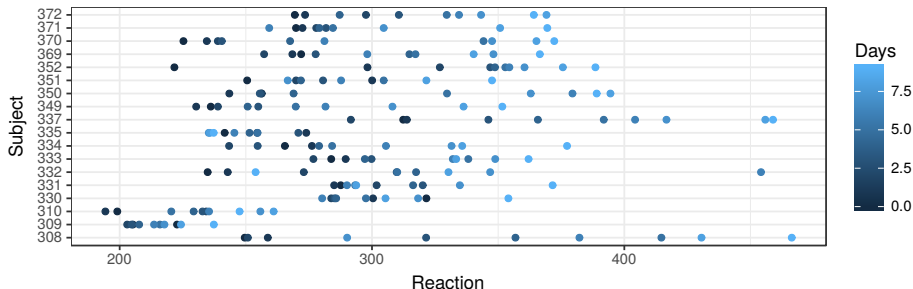
Как меняется время реакции разных субъектов?

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



Как меняется время реакции разных субъектов?

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



- ▶ Видно, что у разных субъектов время реакции различается. Есть быстрые, есть медленные, кого-то недосып стимулирует. Сама по себе межиндивидуальная изменчивость нас не интересует, но ее 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).

The Bad. Не учитываем группирующий фактор.

$$\text{Reaction}_i = \beta_0 + \beta_1 \text{Days}_i + \varepsilon_i$$

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

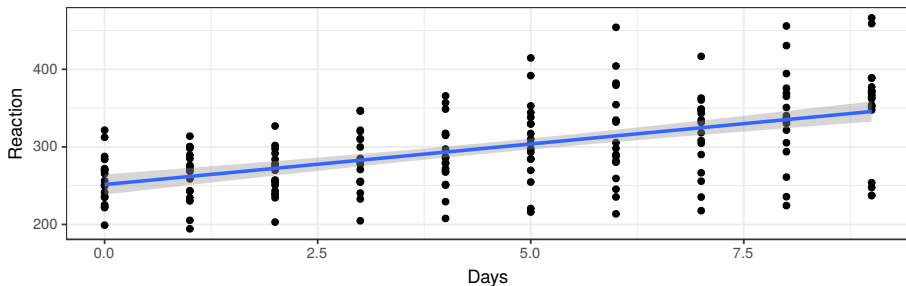
$i = 1, 2, \dots, 180$ – общее число наблюдений

В матричном виде

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

```
Wrong1 <- lm(Reaction ~ Days, data = sl)
```

График этой модели



The Bad. Не учитываем группирующий фактор.

summary(Wrong1)

```
#
# Call:
# lm(formula = Reaction ~ Days, data = sl)
#
# 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
#
# Residual standard error: 47.71 on 178 degrees of freedom
# Multiple R-squared:  0.2865, Adjusted R-squared:  0.2825
# F-statistic: 71.46 on 1 and 178 DF, p-value: 9.894e-15
```

The Bad. Не учитываем группирующий фактор.

summary(Wrong1)

```
#
# Call:
# lm(formula = Reaction ~ Days, data = sl)
#
# 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
#
# Residual standard error: 47.71 on 178 degrees of freedom
# Multiple R-squared:  0.2865, Adjusted R-squared:  0.2825
# F-statistic: 71.46 on 1 and 178 DF, p-value: 9.894e-15
```

- ▶ Если мы не учитываем группирующий фактор, увеличивается вероятность ошибок I рода. Все будет казаться “очень достоверно” из-за низких стандартных ошибок. Но поскольку в этом случае условие независимости нарушено — **все не так, как кажется.**



The Ugly. Группирующий фактор как фиксированный.

$$Reaction_{ij} = \beta_0 + \beta_1 Days_j + \beta_2 Subject_{i=2} + \dots + \beta_{18} Subject_{i=18} + \varepsilon_{ij}$$

$\varepsilon_{ij} \sim N(0, \sigma^2)$ - остатки от регрессии

$i = 1, 2, \dots, 18$ - субъект

$j = 1, 2, \dots, 10$ - день

В матричном виде

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

```
Wrong2 <- lm(Reaction ~ Days + Subject, data = sl)
```

The Ugly. Группирующий фактор как фиксированный.

$$\text{Reaction}_{ij} = \beta_0 + \beta_1 \text{Days}_j + \beta_2 \text{Subject}_{i=2} + \dots + \beta_{18} \text{Subject}_{i=18} + \varepsilon_{ij}$$

$\varepsilon_{ij} \sim N(0, \sigma^2)$ - остатки от регрессии

$i = 1, 2, \dots, 18$ - субъект

$j = 1, 2, \dots, 10$ - день

В матричном виде

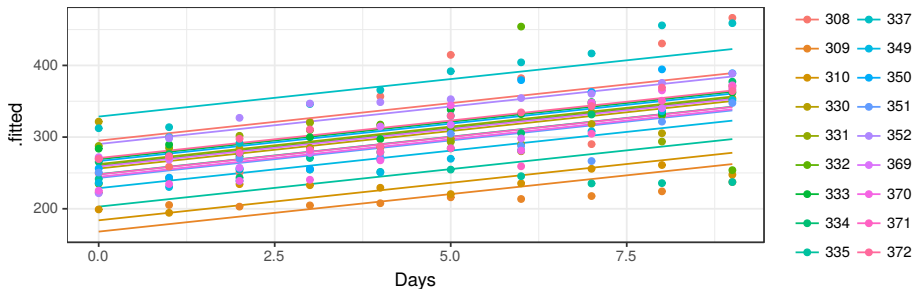
$$\mathbf{Reaction} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

```
Wrong2 <- lm(Reaction ~ Days + Subject, data = sl)
```

Если мы учитываем группирующий фактор как обычно (как **фиксированный фактор**), придется оценивать слишком много параметров (18 для уровней группирующего фактора, 1 для Days, σ — всего 20). При этом у нас всего 180 наблюдений. Чтобы получить удовлетворительную мощность, нужно минимум 10–20 наблюдений на каждый параметр (Harrell, 2013) — у нас 9.

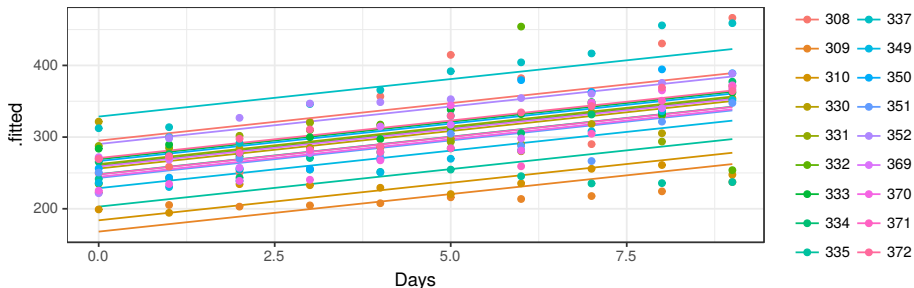
The Ugly. Что нам делать с этим множеством прямых?

```
Wrong2_diag <- fortify(Wrong2)
ggplot(Wrong2_diag, 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))
```



The Ugly. Что нам делать с этим множеством прямых?

```
Wrong2_diag <- fortify(Wrong2)
ggplot(Wrong2_diag, 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 этой модели, где субъект — это фиксированный фактор, для каждого субъекта есть “поправка” для значения свободного члена в уравнении регрессии. В результате универсальность модели теряется: предсказания можно сделать только на индивидуальном уровне — с учетом субъекта



Фиксированные и случайные факторы



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**, а модели с такими факторами называются **смешанными моделями**:

- ▶ Общие смешанные модели (general linear mixed models) — нормальное распределение зависимой переменной
- ▶ Обобщенные смешанные модели (generalized linear mixed models) — другие формы распределений зависимой переменной

Фиксированные и случайные факторы

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 $H_0 : \mu_1 = \mu_2 = \dots = \mu_i = \mu$	о дисперсии отклика между уровнями фактора $H_0 : \sigma_{rand.fact.}^2 = 0$
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 много уровней фактора — не менее 5

Примеры фиксированных и случайных факторов

Фиксированные факторы

- ▶ Пол
- ▶ Низина/вершина
- ▶ Илистый/песчаный грунт
- ▶ Тень/свет
- ▶ Опыт/контроль

Случайные факторы

- ▶ Субъект, особь или площадка (если есть несколько измерений)
- ▶ Выводок (птенцы из одного выводка имеют право быть похожими)
- ▶ Блок, делянка на участке
- ▶ Аквариум в лаб. эксперименте

Задание 1

Какого типа эти факторы? Поясните ваш выбор.

- ▶ Несколько произвольно выбранных градаций плотности моллюсков в полевом эксперименте, где плотностью манипулировали.
- ▶ Фактор размер червяка (маленький, средний, большой) в выборке червей.
- ▶ Деление губы Чупа на зоны с разной степенью распреснения.

Смешанные линейные модели

Смешанная линейная модель в общем виде

$$\mathbf{Y}_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + \mathbf{Z}_i \cdot \mathbf{b}_i + \epsilon_i$$

$\mathbf{b}_i \sim N(0, \mathbf{D})$ — случайные эффекты нормально распределены со средним 0 и матрицей ковариаций \mathbf{D} (дисперсией σ_b^2)

$\epsilon_i \sim N(0, \Sigma)$ — остатки модели нормально распределены со средним 0 и матрицей ковариаций Σ_i (дисперсией σ^2)

$\mathbf{X}_i \cdot \boldsymbol{\beta}$ — фиксированная часть модели

$\mathbf{Z}_i \cdot \mathbf{b}_i$ — случайная часть модели

В примере модель со случайным отрезком можно записать так:

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

$b_i \sim N(0, \sigma_b^2)$ — случайный эффект субъекта (intercept)

$\varepsilon_{ij} \sim N(0, \sigma^2)$ — остатки модели

$i = 1, 2, \dots, 18$ — субъекты

$j = 1, 2, \dots, 10$ — дни

В примере модель со случайным отрезком можно записать так:

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

$b_i \sim N(0, \sigma_b^2)$ — случайный эффект субъекта (intercept)

$\varepsilon_{ij} \sim N(0, \sigma^2)$ — остатки модели

$i = 1, 2, \dots, 18$ — субъекты

$j = 1, 2, \dots, 10$ — дни

Для каждого субъекта i в матричном виде это записывается так:

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

В примере модель со случайным отрезком можно записать так:

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

$b_i \sim N(0, \sigma_b^2)$ — случайный эффект субъекта (intercept)

$\varepsilon_{ij} \sim N(0, \sigma^2)$ — остатки модели

$i = 1, 2, \dots, 18$ — субъекты

$j = 1, 2, \dots, 10$ — дни

Для каждого субъекта i в матричном виде это записывается так:

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

что можно записать сокращенно так:

$$\mathbf{Reaction}_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + \mathbf{Z}_i \cdot \mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

Теперь разберемся с допущениями модели

$$\text{Reaction}_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + \mathbf{Z}_i \cdot \mathbf{b}_i + \epsilon_i$$

$\mathbf{b}_i \sim N(0, \mathbf{D})$ - случайные эффекты b_i нормально распределены со средним 0 и матрицей ковариаций \mathbf{D}

$\epsilon_i \sim N(0, \Sigma_i)$ - остатки модели нормально распределены со средним 0 и матрицей ковариаций Σ_i

Теперь разберемся с допущениями модели

$$\text{Reaction}_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + \mathbf{Z}_i \cdot \mathbf{b}_i + \varepsilon_i$$

$\mathbf{b}_i \sim N(0, \mathbf{D})$ - случайные эффекты b_i нормально распределены со средним 0 и матрицей ковариаций \mathbf{D}

$\varepsilon_i \sim N(0, \Sigma_i)$ - остатки модели нормально распределены со средним 0 и матрицей ковариаций Σ_i

Матрица ковариаций остатков для каждого субъекта выглядит так:

$$\Sigma_i = \sigma^2 \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Теперь разберемся с допущениями модели

$$\text{Reaction}_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + \mathbf{Z}_i \cdot \mathbf{b}_i + \varepsilon_i$$

$\mathbf{b}_i \sim N(0, \mathbf{D})$ - случайные эффекты b_i нормально распределены со средним 0 и матрицей ковариаций \mathbf{D}

$\varepsilon_i \sim N(0, \Sigma_i)$ - остатки модели нормально распределены со средним 0 и матрицей ковариаций Σ_i

Матрица ковариаций остатков для каждого субъекта выглядит так:

$$\Sigma_i = \sigma^2 \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Т.е. остатки независимы друг от друга (вне диагонали стоят нули, т.е. ковариация разных остатков 0).

В то же время, отдельные значения переменной-отклика \mathbf{Y}_i уже не будут независимы друг от друга при добавлении случайных эффектов - см. ниже



Матрица ковариаций переменной-отклика

$$\mathbf{Reaction}_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + \mathbf{Z}_i \cdot \mathbf{b}_i + \varepsilon_i$$

$$\mathbf{b}_i \sim N(0, \mathbf{D})$$

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

Можно показать, что переменная-отклик \mathbf{Y}_i нормально распределена

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

Матрица ковариаций переменной-отклика

$$\mathbf{Reaction}_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + \mathbf{Z}_i \cdot \mathbf{b}_i + \varepsilon_i$$

$$\mathbf{b}_i \sim N(0, \mathbf{D})$$

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

Можно показать, что переменная-отклик \mathbf{Y}_i нормально распределена

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

Матрица ковариаций переменной-отклика:

$$\mathbf{V}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \Sigma_i$$

где \mathbf{D} — матрица ковариаций случайных эффектов.

Т.е. **добавление случайных эффектов приводит к изменению ковариационной матрицы \mathbf{V}_i**

Кстати, $\mathbf{Z}_i \mathbf{D} \mathbf{Z}_i'$ называется преобразование Холецкого (Cholesky decomposition)

Добавление случайных эффектов приводит к изменению ковариационной матрицы

$$\mathbf{v}_i = \mathbf{z}_i \mathbf{D} \mathbf{z}_i' + \Sigma_i$$

Для простейшей смешанной модели со случайным отрезком:

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

Индукционная корреляция - следствие включения в модель случайных эффектов

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

Индукционная корреляция - следствие включения в модель случайных эффектов

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

σ_b^2 — ковариация между наблюдениями одного субъекта

$\sigma^2 + \sigma_b^2$ — дисперсия

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

Индукционная корреляция - следствие включения в модель случайных эффектов

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

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

Способ измерить, насколько коррелируют друг с другом наблюдения из одной и той же группы случайного фактора. Если он высок, то можно брать меньше проб в группе (и больше групп, если нужно)

Подбор смешанных моделей в R

Подбор смешанных моделей в R

Самые популярные пакеты — `nlme` (старый, иногда медленный, стабильный, хорошо документированный) и `lme4` (новый, быстрый, не такой стабильный, хуже документированный). Есть много других.

Функция	<code>lme()</code> из <code>nlme</code>	<code>lmer()</code> из <code>lme4</code>	<code>glmer()</code> из <code>lme4</code>	<code>glmmPQL()</code> из <code>MASS</code>
Распределение отклика	нормальное	нормальное	биномиальное, пуассоновское, гамма, (+ квази)	биномиальное, пуассоновское, гамма, (+ квази), отр. биномиальное
Метод оценивания	ML, REML	ML, REML	ML, REML	PQL
Гетерогенность дисперсий	+	-	-	-
Корреляционные структуры	+	-	-	+
Доверительная вероятность (p-value)	+	-	-	+

Фиксированная часть модели задается обычной двухсторонней формулой

$$Y \sim 1 + X1 + \dots + Xn$$

Случайная часть модели - односторонняя формула. До вертикальной черты — перечислены факторы, влияющие на случайный угол наклона. После вертикальной черты — факторы, влияющие на случайный intercept.

$$\sim 1 + X1 + \dots + Xn \mid A$$

Вложенные друг в друга факторы указываются от крупного к мелкому через “/”

$$\sim 1 + X1 + \dots + Xn \mid A/B/C$$

Детали синтаксиса разных функций отличаются (см. следующий слайд с примерами формул)

Синтаксис некоторых смешанных моделей

Факторы	lme() из nlme	lmer() из lme4
A – случ. intercept	<code>lme(fixed=Y~1,random=~1 A, data=dt)</code>	<code>lmer(Y~1+(1 A), data=dt)</code>
A – случ. intercept, X – фикс.	<code>lme(fixed=Y~X,random=~1 A, data=dt)</code>	<code>lmer(Y~X+(1 A), data=dt)</code>
A – случ. intercept, X – случ. угол накл.	<code>lme(fixed=Y~X,random=~1+X A, data=dt)</code>	<code>lmer(Y~X+(1+X A), data=dt)</code>
A и B – случ. intercept, A и B независимы (crossed effects), X – фикс.		<code>lmer(Y~X+(1 A)+(1 B), data=dt)</code>
A и B – случ. intercept, B вложен в A (nested effects), уровни B повт. в группах по фактору A, X – фикс.	<code>lme(fixed=Y~X,random=~1 A/B, data=dt)</code>	<code>lmer(Y~X+(1 A/B), data=dt)</code> <code>lmer(Y~X+(1 A)+(1 A:B), data=dt)</code>
A и B – случ. intercept, B вложен в A (nested random effects), все уровни B уникальны, X – фикс.	<code>lme(fixed=Y~X,random=~1 A/B, data=dt)</code>	<code>lmer(Y~X+(1 A)+(1 B), data=dt)</code>

Смешанные модели со случайным отрезком в R

Подберем модель со случайным отрезком с помощью `lme()` из пакета `nlme`.

Функция `lme()` из пакета `nlme` нам понадобится на следующем занятии, поэтому нужно освоить ее синтаксис.

```
# выгружаем lme4, чтобы не было конфликтов с nlme  
detach(name = "package:lme4")  
library(nlme)  
M1 <- lme(Reaction ~ Days, random = ~ 1 | Subject, data = sl)
```

Что дальше?



Анализ остатков

График остатков от предсказанных значений

```
M1_diag <- data.frame(sl,  
  .resid = resid(M1, type = "pearson"),  
  .fitted <- fitted(M1))  
gg_resid <- ggplot(M1_diag, aes(y = .resid)) +  
  guides(colour = guide_legend(ncol = 2))  
gg_resid + geom_point(aes(x = .fitted, colour = Subject))
```

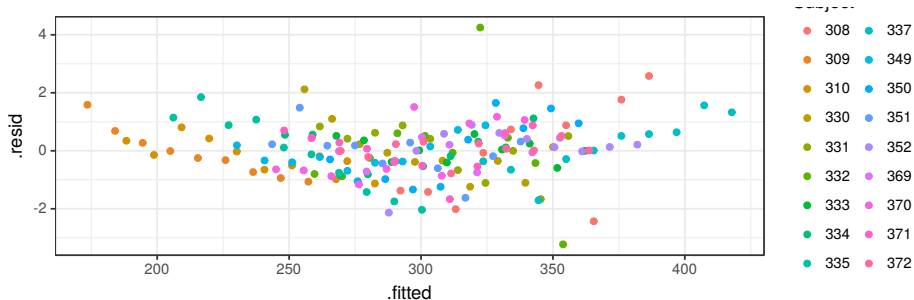
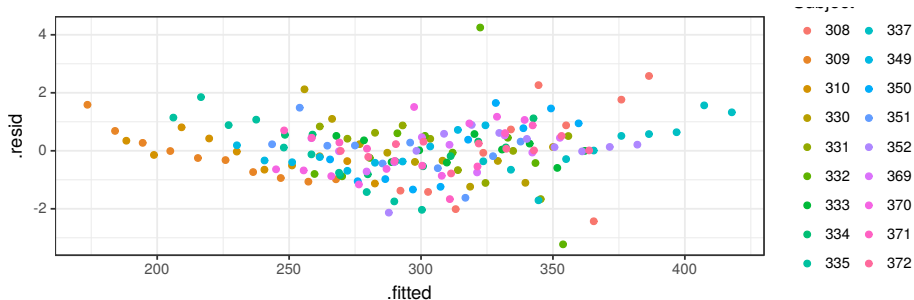


График остатков от предсказанных значений

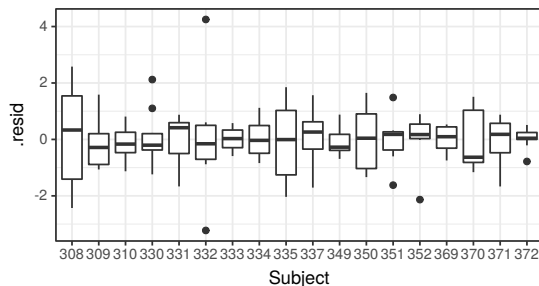
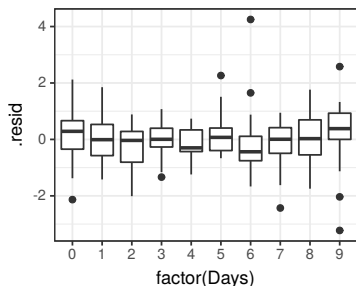
```
M1_diag <- data.frame(sl,  
  .resid = resid(M1, type = "pearson"),  
  .fitted <- fitted(M1))  
gg_resid <- ggplot(M1_diag, aes(y = .resid)) +  
  guides(colour = guide_legend(ncol = 2))  
gg_resid + geom_point(aes(x = .fitted, colour = Subject))
```



- Есть большие остатки, гетерогенность дисперсий

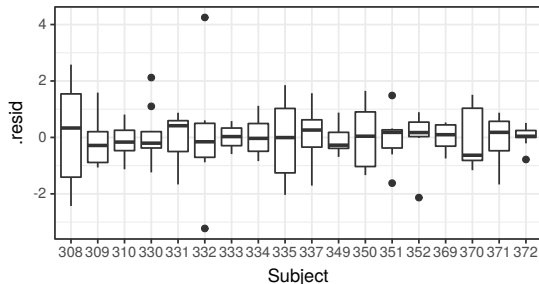
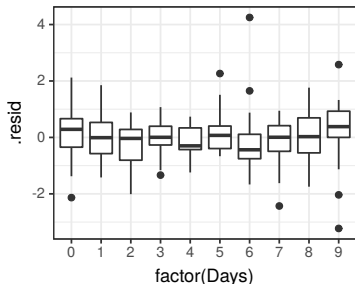
Графики остатков от ковариат в модели и не в модели

```
library(gridExtra)
grid.arrange(gg_resid + geom_boxplot(aes(x = factor(Days))),
             gg_resid + geom_boxplot(aes(x = Subject)),
             ncol = 2, widths = c(0.4, 0.6))
```



Графики остатков от ковариат в модели и не в модели

```
library(gridExtra)
grid.arrange(gg_resid + geom_boxplot(aes(x = factor(Days))),
             gg_resid + geom_boxplot(aes(x = Subject)),
             ncol = 2, widths = c(0.4, 0.6))
```



- ▶ Большие остатки у наблюдений для 332 субъекта
- ▶ Гетерогенность дисперсий
- ▶ Пока оставим все как есть

Тестирование гипотез в смешанных моделях

Способы тестирования влияния факторов в смешанных моделях

Достаточно **одного** из этих равноправных вариантов.

Важно, каким именно способом (ML или REML) подобрана модель.

(а) t-(или -z) тесты — приблизительный результат (REML)

(б) F-тест — приблизительный результат (REML)

(в) Попарное сравнение вложенных моделей при помощи тестов отношения правдоподобий (ML)

(г) Сравнение моделей по AIC (ML)

(a) t-(или -z) тесты (REML)

- ▶ `summary(model)`
- ▶ Подходит для непрерывных переменных или факторов с 2 уровнями.
- ▶ Дает приблизительный результат, лучше так не делать.

```
summary(M1)
```

```
# Linear mixed-effects model fit by REML
# Data: sl
#      AIC      BIC    logLik
# 1794.465 1807.192 -893.2325
#
# Random effects:
# Formula: ~1 | Subject
#      (Intercept) Residual
# StdDev:      37.12383 30.99123
#
# Fixed effects: Reaction ~ Days
#               Value Std.Error DF t-value p-value
# (Intercept) 251.40510  9.746716 161 25.79383      0
# Days        10.46729  0.804221 161 13.01543      0
# Correlation:
#      (Intr)
# Days -0.371
#
# Standardized Within-Group Residuals:
#      Min      Q1      Med      Q3      Max
# -3.2256707 -0.5528788  0.0108521  0.5187971  4.2506162
#
```

(6) F-тест (REML)

- ▶ `anova()`
- ▶ Приблизительный результат, лучше так не делать.
- ▶ Последовательное тестирование гипотез (Type I SS) — будьте внимательны при интерпретации

```
library(car)
anova(M1, test = "F")
```

#	numDF	denDF	F-value	p-value
# (Intercept)	1	161	1087.9793	<.0001
# Days	1	161	169.4014	<.0001

(6) F-тест (REML)

- ▶ `anova()`
- ▶ Приблизительный результат, лучше так не делать.
- ▶ Последовательное тестирование гипотез (Type I SS) — будьте внимательны при интерпретации

```
library(car)
anova(M1, test = "F")
```

#	numDF	denDF	F-value	p-value
# (Intercept)	1	161	1087.9793	<.0001
# Days	1	161	169.4014	<.0001

- ▶ Время реакции зависит от продолжительности бессонницы ($F_{1,161} = 169$, $p < 0.01$)

(в) Попарное сравнение вложенных моделей при помощи тестов отношения правдоподобий (ML)

Дает более точные выводы, чем F и t(z)

Обязательно method = "ML", а не "REML"

```
M1.ml <- lme(Reaction ~ Days, random = ~1|Subject, data = sl, method = "ML")
M2.ml <- lme(Reaction ~ 1, random = ~1 | Subject, data = sl, method = "ML")
```

Любой из этих вариантов:

- ▶ `anova(model1, model2)`
- ▶ `drop1()`
- ▶ `Anova()` из пакета `car` — Type II, III SS, не приводится значение отношения правдоподобий

```
anova(M1.ml, M2.ml)
```

#	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
# M1.ml	1	4	1802.079	1814.851	-897.0393			
# M2.ml	2	3	1916.541	1926.120	-955.2705	1 vs 2	116.4624	<.0001

(в) Попарное сравнение вложенных моделей при помощи тестов отношения правдоподобий (ML)

Дает более точные выводы, чем F и t(z)

Обязательно method = "ML", а не "REML"

```
M1.ml <- lme(Reaction ~ Days, random = ~1|Subject, data = sl, method = "ML")
M2.ml <- lme(Reaction ~ 1, random = ~1 | Subject, data = sl, method = "ML")
```

Любой из этих вариантов:

- ▶ `anova(model1, model2)`
- ▶ `drop1()`
- ▶ `Anova()` из пакета `car` — Type II, III SS, не приводится значение отношения правдоподобий

```
anova(M1.ml, M2.ml)
```

#	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
#	M1.ml	1 4	1802.079	1814.851	-897.0393			
#	M2.ml	2 3	1916.541	1926.120	-955.2705	1 vs 2	116.4624	<.0001

- ▶ Время реакции меняется в зависимости от продолжительности бессонницы ($L = 116$, $df = 1$, $p < 0.01$)



(г) Сравнение моделей по AIC (ML)

Обязательно `method = "ML"`, а не `"REML"`

```
AIC(M1.ml, M2.ml)
```

#		df	AIC
#	M1.ml	4	1802.079
#	M2.ml	3	1916.541

(г) Сравнение моделей по AIC (ML)

Обязательно method = "ML", а не "REML"

```
AIC(M1.ml, M2.ml)
```

#		df	AIC
#	M1.ml	4	1802.079
#	M2.ml	3	1916.541

- ▶ Продолжительность бессонницы влияет на время реакции (AIC)

Подбор оптимальной модели и проверка условий применимости

Подбор оптимальной модели и проверка условий применимости

Если вы решили подбирать оптимальную модель и выкидывать какие-то предикторы, то вам нужно будет сделать анализ остатков финальной модели.

В нашем случае модель не изменилась, поэтому данный этап выпадает из анализа

Представление результатов

Представление результатов

REML оценка параметров более точна (оценка случайных факторов)

Для представления результатов лучше использовать модель, подобранную при помощи Restricted Maximum Likelihood.

```
M1_fin <- lme(Reaction ~ Days, random = ~1|Subject, data = sl,  
             method = "REML")
```

В данном случае, этот шаг избыточен, т.к. lme использует REML по-умолчанию, и поэтому сейчас нам не нужно было ничего менять.

Но lmer использует ML, и тогда точно нужно переподобрать финальную модель при помощи REML.

Уравнение модели

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

$b_i \sim N(0, 31^2)$ — случайный эффект субъекта

$\varepsilon_{ij} \sim N(0, 37.1^2)$ — остатки модели

$i = 1, 2, \dots, 18$ — субъекты

$j = 1, 2, \dots, 10$ — дни

```
fixef(M1_fin)    # Фиксированные эффекты
```

```
# (Intercept)      Days
#    251.40510     10.46729
```

```
VarCorr(M1_fin)  # Случайные эффекты
```

```
# Subject = pdLogChol(1)
#              Variance StdDev
# (Intercept) 1378.1785 37.12383
# Residual    960.4566 30.99123
```

Внутриклассовая корреляция

$$\sigma_{effect}^2 / (\sigma_{effect}^2 + \sigma^2)$$

Внутриклассовая корреляция

```
37.12383^2 / (37.12383^2 + 30.99123^2)
```

```
# [1] 0.589309
```

```
M1_fin
```

В результатах

Random effects:

Formula: ~1 | Subject

(Intercept) Residual

StdDev: 37.12383 30.99123

Внутриклассовая корреляция

$$\sigma_{effect}^2 / (\sigma_{effect}^2 + \sigma^2)$$

Внутриклассовая корреляция

```
37.12383^2 / (37.12383^2 + 30.99123^2)
```

```
# [1] 0.589309
```

```
M1_fin
```

В результатах

Random effects:

Formula: ~1 | Subject

(Intercept) Residual

StdDev: 37.12383 30.99123

- ▶ Значения времени реакции одного субъекта похожи. Высокая внутриклассовая корреляция показывает, что эффект субъекта нельзя игнорировать в анализе.



Данные для графика предсказаний фиксированной части модели

```
# Исходные данные
library(plyr)
NewData_M1 <- ddply(
  sl, .(Subject), summarise,
  Days = seq(min(Days), max(Days), length = 10)
)

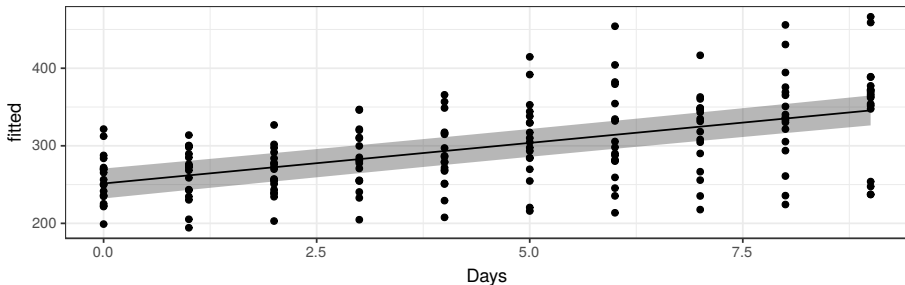
# Предсказанные значения при помощи predict()
# level = 0 - для фиксированных эффектов (т.е. без учета субъекта)
NewData_M1$fitted <- predict(M1_fin, NewData_M1, level = 0)

# Предсказанные значения при помощи матриц
X <- model.matrix(~ Days, data = NewData_M1)
betas <- fixef(M1_fin)
NewData_M1$fitted <- X %*% betas

# Стандартные ошибки и дов. интервалы
NewData_M1$se <- sqrt( diag(X %*% vcov(M1_fin) %*% t(X)) )
NewData_M1$lwr <- NewData_M1$fitted - 1.98 * NewData_M1$se
NewData_M1$upr <- NewData_M1$fitted + 1.98 * NewData_M1$se
```

График предсказаний фиксированной части модели

```
ggplot(data = NewData_M1, aes(x = Days, y = fitted)) +  
  geom_ribbon(alpha = 0.35, aes(ymin = lwr, ymax = upr)) +  
  geom_line() +  
  geom_point(data = sl, aes(x = Days, y = Reaction))
```



Данные для графика предсказаний для индивидуальных уровней случайного фактора

Если вам любопытно, куда делась информация о разных субъектах, то вот она...

Можно получить предсказания для каждого субъекта

$$\beta_0 + \beta_1 \cdot Days_{ij} + b_i$$

```
NewData_M1$fit_subj <- predict(M1_fin, NewData_M1, level = 1)
# или то же самое при помощи матриц
# случайные эффекты для каждого субъекта
# это датафрейм с одним столбцом
rand <- ranef(M1_fin)
# "разворачиваем" для каждой строки данных
all_rand <- rand[as.numeric(NewData_M1$Subject), 1]
# прибавляем случайные эффекты к предсказаниям фикс. части
NewData_M1$fit_subj <- X %*% betas + all_rand
```

График предсказаний для индивидуальных уровней случайного фактора

```
ggplot(NewData_M1, aes(x = Days, y = fit_subj, group = Subject)) +  
  geom_ribbon(alpha = 0.5, aes(fill = Subject, ymin = fit_subj - 1.98*se,  
    ymax = fit_subj + 1.98*se)) +  
  geom_line() +  
  geom_point(data = sl, aes(x = Days, y = Reaction)) +  
  guides(fill = guide_legend(ncol = 2))
```

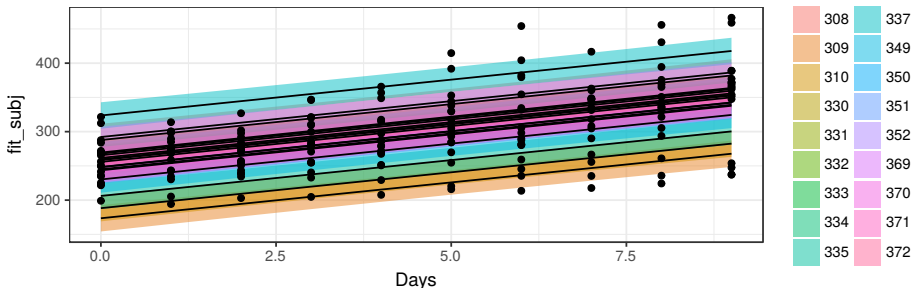
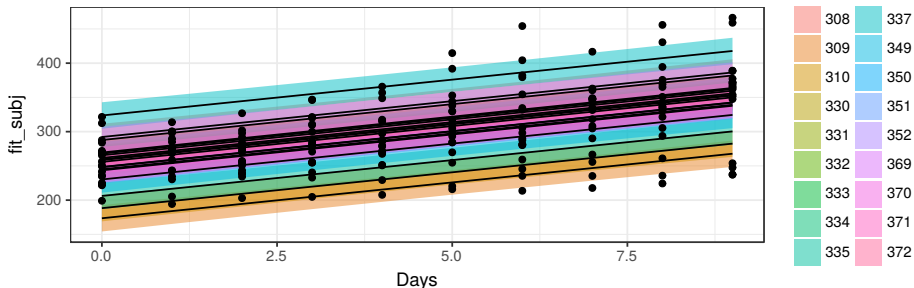


График предсказаний для индивидуальных уровней случайного фактора

```
ggplot(NewData_M1, aes(x = Days, y = fit_subj, group = Subject)) +  
  geom_ribbon(alpha = 0.5, aes(fill = Subject, ymin = fit_subj - 1.98*se,  
    ymax = fit_subj + 1.98*se)) +  
  geom_line() +  
  geom_point(data = sl, aes(x = Days, y = Reaction)) +  
  guides(fill = guide_legend(ncol = 2))
```



Не факт, что на самом деле время реакции разных субъектов меняется параллельно



Смешанные модели со случайным отрезком и углом наклона в R

Смешанная модель со случайным отрезком и углом наклона

На графике индивидуальных эффектов было видно, что измерения для разных субъектов, возможно, идут непараллельными линиями. 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

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

$b_i \sim N(0, \sigma_b^2)$ — случайный интерсепт для субъекта

$c_{ij} \sim N(0, \sigma_c^2)$ — случайный угол наклона для субъекта

$\varepsilon_{ij} \sim N(0, \sigma^2)$ — остатки модели

$i = 1, 2, \dots, 18$ — субъекты

$j = 1, 2, \dots, 10$ — дни

Дальнейшие действия по прежнему плану:

- ▶ Подбираем модель
- ▶ Анализ остатков
- ▶ Проверка влияния факторов + подбор оптимальной модели
- ▶ Анализ остатков финальной модели
- ▶ Подбор финальной модели при помощи REML
- ▶ Описание результатов
- ▶ Визуализация предсказаний

Смешанная модель со случайным отрезком и углом наклона

Формат записи формулы для случайных эффектов в `lme()`:

`random = ~ 1 + Угол наклона | Интерсепт`

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

Задание 2

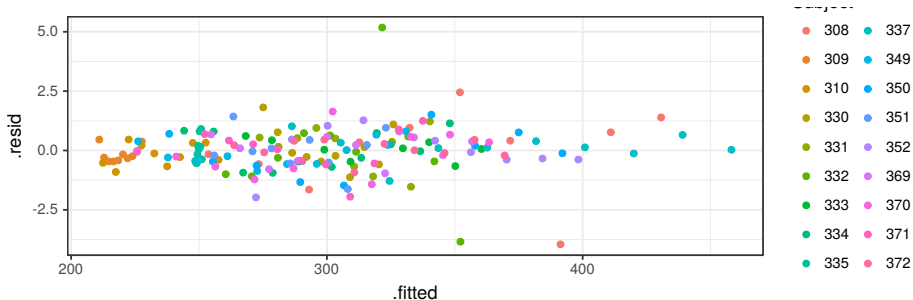
Проверьте получившуюся модель MS1

Сделайте самостоятельно:

- ▶ Анализ остатков
- ▶ Проверку влияния факторов + подбор оптимальной модели
- ▶ Визуализацию предсказаний

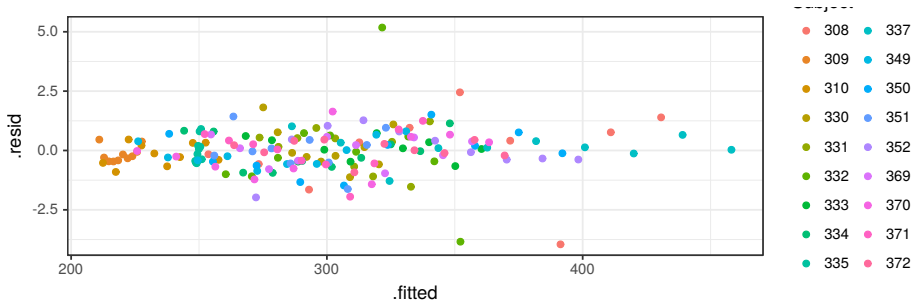
Решение: График остатков от предсказанных значений

```
MS1_diag <- data.frame(sl,  
  .resid = resid(MS1, type = "pearson"),  
  .fitted <- fitted(MS1))  
gg_resid_1 <- ggplot(MS1_diag, aes(y = .resid)) +  
  guides(colour = guide_legend(ncol = 2))  
gg_resid_1 + geom_point(aes(x = .fitted, colour = Subject))
```



Решение: График остатков от предсказанных значений

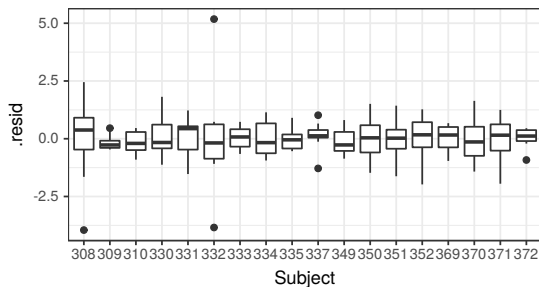
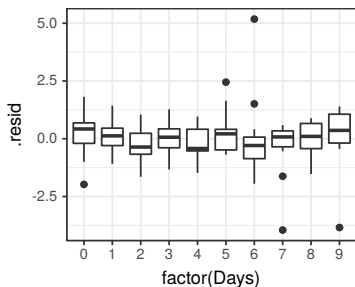
```
MS1_diag <- data.frame(sl,  
  .resid = resid(MS1, type = "pearson"),  
  .fitted <- fitted(MS1))  
gg_resid_1 <- ggplot(MS1_diag, aes(y = .resid)) +  
  guides(colour = guide_legend(ncol = 2))  
gg_resid_1 + geom_point(aes(x = .fitted, colour = Subject))
```



- Есть большие остатки, гетерогенность дисперсий не выражена

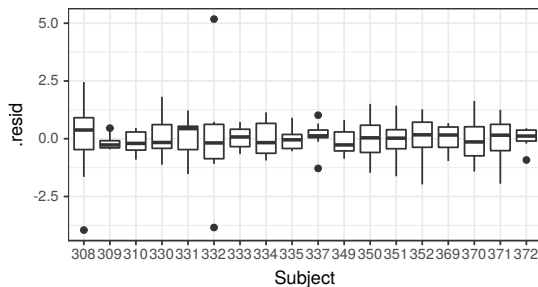
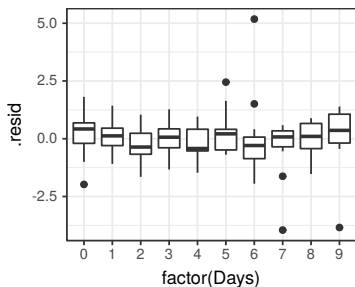
Решение: Графики остатков от ковариат в модели и не в модели

```
grid.arrange(gg_resid_1 + geom_boxplot(aes(x = factor(Days))),  
             gg_resid_1 + geom_boxplot(aes(x = Subject)),  
             ncol = 2, widths = c(0.4, 0.6))
```



Решение: Графики остатков от ковариат в модели и не в модели

```
grid.arrange(gg_resid_1 + geom_boxplot(aes(x = factor(Days))),  
             gg_resid_1 + geom_boxplot(aes(x = Subject)),  
             ncol = 2, widths = c(0.4, 0.6))
```



- ▶ Большие остатки у наблюдений 332 субъекта
- ▶ Гетерогенность дисперсий уже не так сильно выражена, как в прошлый раз.

Решение: Проверка влияния факторов

Тестируем значимость влияния продолжительности бессонницы. Сделаем это при помощи теста отношения правдоподобий.

```
MS1.ml <- lme(Reaction ~ Days, random = ~1 + Days|Subject, data = sl,  
             method = "ML")  
MS2.ml <- update(MS1.ml, . ~ . - Days)  
anova(MS1.ml, MS2.ml)
```

#	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
# MS1.ml	1	6	1763.939	1783.097	-875.9697			
# MS2.ml	2	5	1785.476	1801.441	-887.7379	1 vs 2	23.53654	<.0001

Решение: Проверка влияния факторов

Тестируем значимость влияния продолжительности бессонницы. Сделаем это при помощи теста отношения правдоподобий.

```
MS1.ml <- lme(Reaction ~ Days, random = ~1 + Days|Subject, data = sl,  
             method = "ML")  
MS2.ml <- update(MS1.ml, . ~ . - Days)  
anova(MS1.ml, MS2.ml)
```

#	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
#	MS1.ml	1	6	1763.939	1783.097	-875.9697		
#	MS2.ml	2	5	1785.476	1801.441	-887.7379	1 vs 2	23.53654 <.0001

- ▶ Время реакции меняется в зависимости от продолжительности бессонницы ($L = 24$, $df = 1$, $p < 0.01$).

Решение: Проверка влияния факторов (случайный интерсепт для субъектов)

Почему мы не тестируем значимость самого фактора Subject?

Потому что этот фактор у нас должен быть в модели по-определению, без обсуждения — из-за того, что у нас такой дизайн эксперимента.

Решение: Проверка влияния факторов (случайный угол наклона для субъектов)

Можем проверить, значимы ли изменения угла наклона для разных субъектов.

Это случайный фактор — используем REML

```
MS1.reml <- lme(Reaction ~ Days, random = ~1 + Days|Subject, data = sl,  
               method = "REML")  
MS3.reml <- update(MS1.reml, random = ~1|Subject)  
anova(MS1.reml, MS3.reml)
```

#	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
# MS1.reml	1	6	1755.628	1774.719	-871.8141			
# MS3.reml	2	4	1794.465	1807.192	-893.2325	1 vs 2	42.83681	<.0001

Решение: Проверка влияния факторов (случайный угол наклона для субъектов)

Можем проверить, значимы ли изменения угла наклона для разных субъектов.

Это случайный фактор — используем REML

```
MS1.reml <- lme(Reaction ~ Days, random = ~1 + Days|Subject, data = sl,  
               method = "REML")  
MS3.reml <- update(MS1.reml, random = ~1|Subject)  
anova(MS1.reml, MS3.reml)
```

#	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
# MS1.reml	1	6	1755.628	1774.719	-871.8141			
# MS3.reml	2	4	1794.465	1807.192	-893.2325	1 vs 2	42.83681	<.0001

- Скорость изменений зависит от субъекта ($L = 43$, $df = 2$, $p < 0.01$)

Решение: Представление результатов

Для представления результатов переподбираем модель заново, используя Restricted Maximum Likelihood.

REML оценка параметров более точна (оценка случайных факторов)

```
MS1_fin <- lme(Reaction ~ Days, random = ~1 + Days|Subject, data = sl,  
              method = "REML")
```

Здесь это избыточный шаг, у нас уже есть такая модель — MS1.reml

Решение: Уравнение модели

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

$b_i \sim N(0, 24.7^2)$ — случайный интерсепт для субъекта

$c_{ij} \sim N(0, 5.9^2)$ — случайный угол наклона для субъекта

$\varepsilon_{ij} \sim N(0, 25.6^2)$ — остатки модели

$i = 1, 2, \dots, 18$ — субъекты

$j = 1, 2, \dots, 10$ — дни

```
fixef(MS1_fin)    # Фиксированные эффекты
```

```
# (Intercept)      Days  
#   251.40510     10.46729
```

```
VarCorr(MS1_fin)  # Случайные эффекты
```

```
# Subject = pdLogChol(1 + Days)  
#           Variance StdDev   Corr  
# (Intercept) 612.0795 24.740241 (Intr)  
# Days        35.0713  5.922103 0.066  
# Residual    654.9424 25.591843
```

Решение: Данные для графика предсказаний фиксированной части модели

Исходные данные

```
NewData_MS1 <- ddp1y(  
  sl, .(Subject), summarise,  
  Days = seq(min(Days), max(Days), length = 10)  
)
```

Предсказанные значения при помощи predict()

level = 0 - для фиксированных эффектов (т.е. без учета субъекта)

```
NewData_MS1$fit <- predict(MS1_fin, NewData_MS1, level = 0)
```

Предсказанные значения при помощи матриц

```
X <- model.matrix(~ Days, data = NewData_MS1)
```

```
betas = fixef(MS1_fin)
```

```
NewData_MS1$fit <- X %*% betas
```

Стандартные ошибки и дов. интервалы

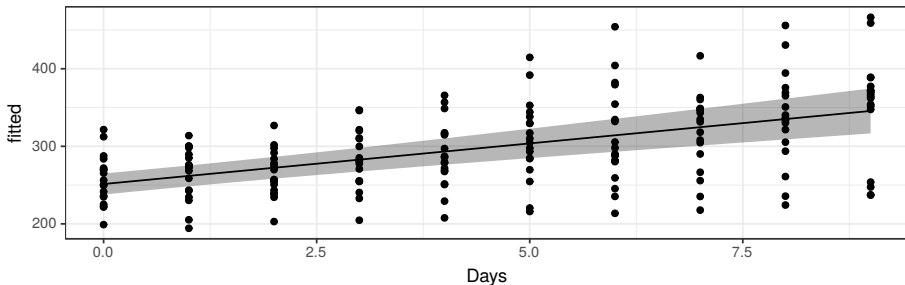
```
NewData_MS1$se <- sqrt(diag(X %*% vcov(MS1_fin) %*% t(X)) )
```

```
NewData_MS1$lwr <- NewData_MS1$fit - 1.98 * NewData_MS1$se
```

```
NewData_MS1$upr <- NewData_MS1$fit + 1.98 * NewData_MS1$se
```

Решение: График предсказаний фиксированной части модели

```
ggplot(data = NewData_MS1, aes(x = Days, y = fitted)) +  
  geom_ribbon(alpha = 0.35, aes(ymin = lwr, ymax = upr)) +  
  geom_line() +  
  geom_point(data = sl, aes(x = Days, y = Reaction))
```



Решение: Данные для графика предсказаний для индивидуальных уровней случайного фактора

Если вам любопытно, куда делась информация о разных субъектах, то вот она...

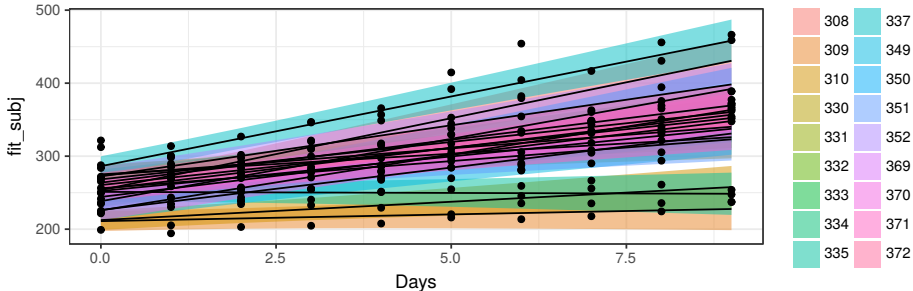
Можно получить предсказания для каждого субъекта

$$\beta_0 + \beta_1 \cdot Days_{ij} + b_i + c_{ij} \cdot Days_{ij}$$

```
NewData_MS1$fit_subj <- predict(MS1_fin, NewData_MS1, level = 1)
# или то же самое при помощи матриц
# случайные эффекты для каждого субъекта
# это датафрейм с двумя столбцами
rand <- ranef(MS1_fin)
# "разворачиваем" для каждой строки данных
all_rand <- rand[as.numeric(NewData_MS1$Subject), ]
# прибавляем случайные эффекты к предсказаниям фикс. части
NewData_MS1$fit_subj <- (betas[1] + all_rand[, 1]) + (betas[2] + all_rand[, 2])
```

Решение: График предсказаний для индивидуальных уровней случайного фактора

```
ggplot(NewData_MS1, aes(x = Days, y = fit_subj, group = Subject)) +  
  geom_ribbon(alpha = 0.5, aes(fill = Subject, ymin = fit_subj - 1.98*se,  
                               ymax = fit_subj + 1.98*se)) +  
  geom_line() +  
  geom_point(data = sl, aes(x = Days, y = Reaction)) +  
  guides(fill = guide_legend(ncol = 2))
```



- ▶ 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.
- ▶ 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.

- ▶ Crawley, M.J. (2007). The R Book (Wiley).
- ▶ 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).