# Смешанные линейные модели

## Линейные модели...

Марина Варфоломеева, Вадим Хайтов

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

# Пример: Как время реакции людей зависит от бессонницы?

Данные из Belenky et al., 2003.
В нулевой день эксперимента всем испытуемым давали поспать нормальное время. Начиная со следующей ночи давали спать по 3 часа.

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

```
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
sapply(sl, function(x) sum(is.na(x)))
```

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

```
# число субъектов
length(unique(sl$Subject))
```

```
# [1] 18
```

# Знакомство с данными (продолжение)

```
# сбалансирован ли объем выборки?
table(sl$Subject)
```

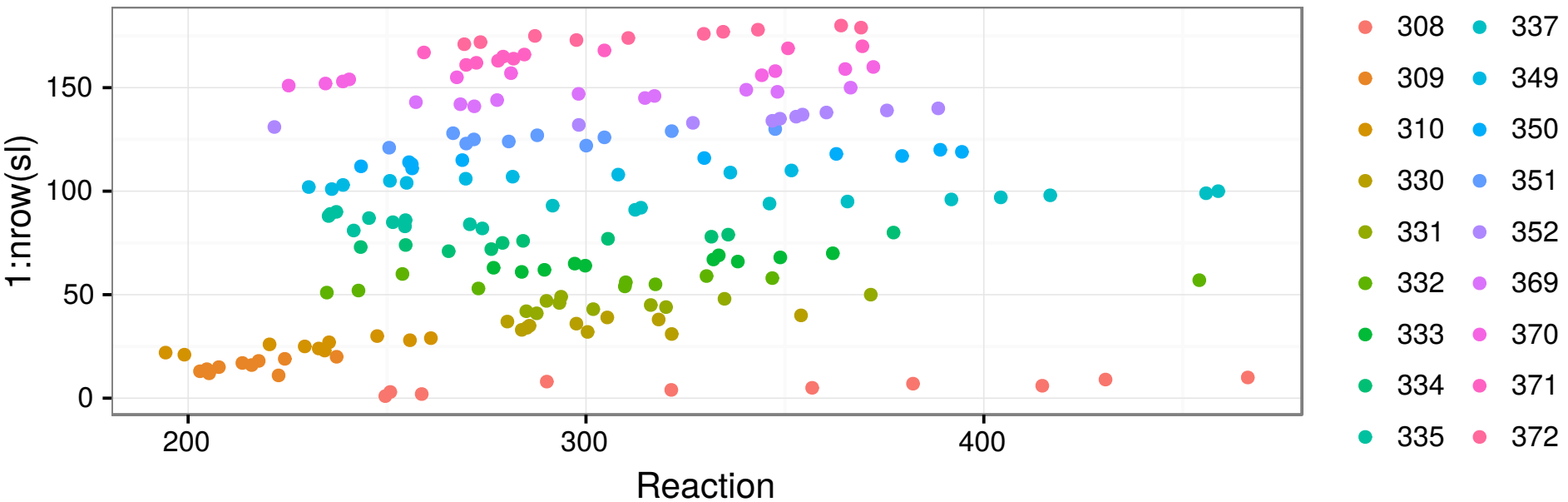
```
#
# 308 309 310 330 331 332 333 334 335 337 349 350 351 352 369 370 371 372
#  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10
```

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

```
#         Days
# Subject 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
#     334 1 1 1 1 1 1 1 1 1 1
#     335 1 1 1 1 1 1 1 1 1 1
#     337 1 1 1 1 1 1 1 1 1 1
#     349 1 1 1 1 1 1 1 1 1 1
#     350 1 1 1 1 1 1 1 1 1 1
```

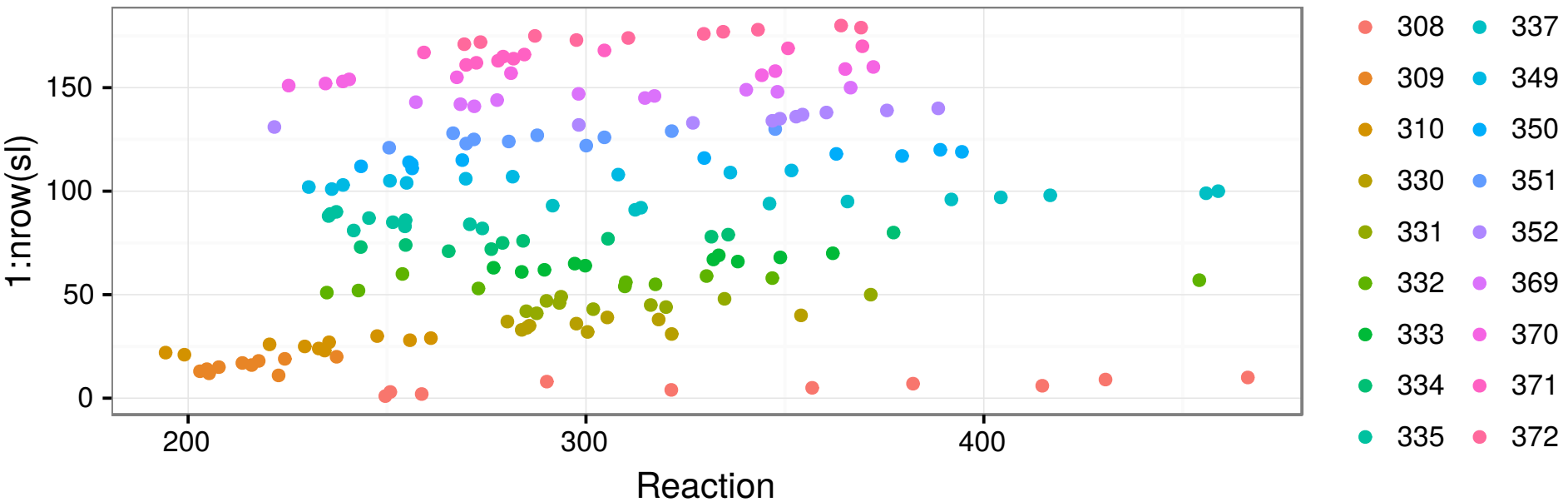
# Есть ли выбросы?

```
library(ggplot2)
theme_set(theme_bw() + theme(legend.key = element_blank()))
update_geom_defaults("point", list(shape = 19))

ggplot(sl, aes(x = Reaction, y = 1:nrow(sl), colour = Subject)) +
  geom_point() + guides(colour = guide_legend(ncol = 2))
```



# Есть ли выбросы?

```
library(ggplot2)
theme_set(theme_bw() + theme(legend.key = element_blank()))
update_geom_defaults("point", list(shape = 19))

ggplot(sl, aes(x = Reaction, y = 1:nrow(sl), colour = Subject)) +
  geom_point() + guides(colour = guide_legend(ncol = 2))
```




- Субъектов с необычным временем реакции нет
- Видно, что у разных субъектов время реакции различается. Есть быстрые, есть медленные. 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`.

# Что делать с разными субъектами?

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

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`.

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` как обычный фиксированный фактор).

# The Bad. Не учитываем группирующий фактор.

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

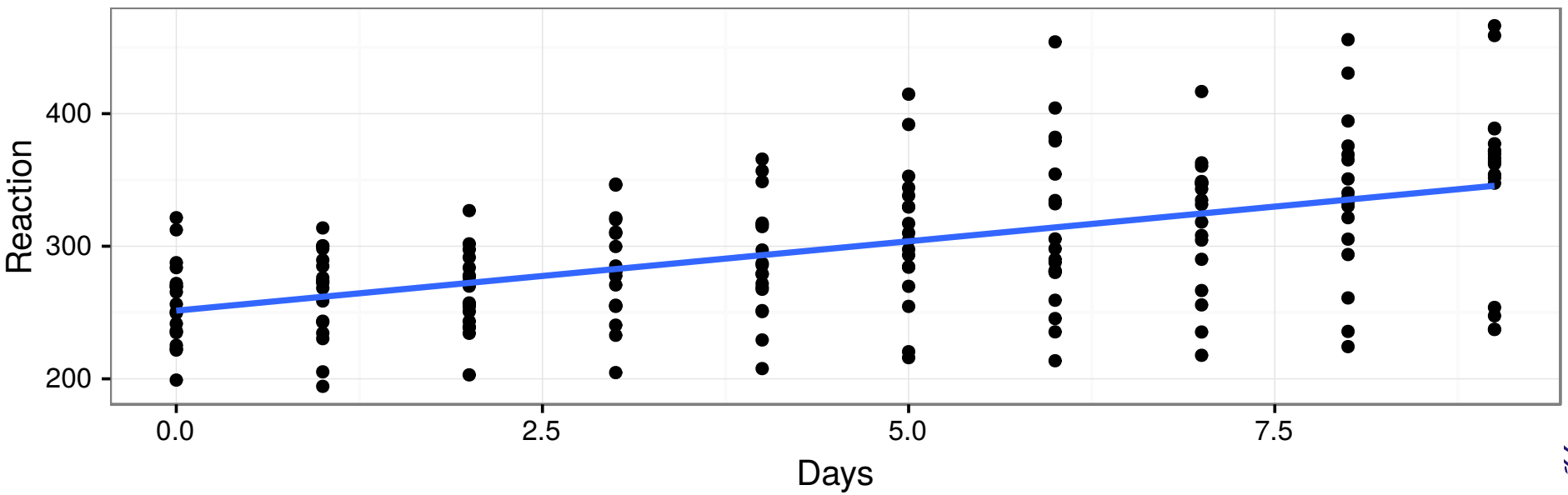
$\varepsilon_i \sim N(0, \sigma^2)$ $i = 1, 2, ..., 180$ – общее число наблюдений

В матричном виде

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

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

График этой модели



# The Bad. Не учитываем группирующий фактор.

```
summary(W1)
```

```
#
# 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(W1)
```

```
#
# 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.

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

$\varepsilon_{ij} \sim N(0, \sigma^2)$ - остатки от регрессии $i = 1, 2, ..., 18$ - субъект $j = 1, 2, ..., 10$ - день

В матричном виде

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

```
W2 <- lm(Reaction ~ Days + Subject, data = sl)
```

# The Ugly. Группирующий фактор как фиксированный.

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

$\varepsilon_{ij} \sim N(0, \sigma^2)$ - остатки от регрессии $i = 1, 2, ..., 18$ - субъект $j = 1, 2, ..., 10$ - день

В матричном виде

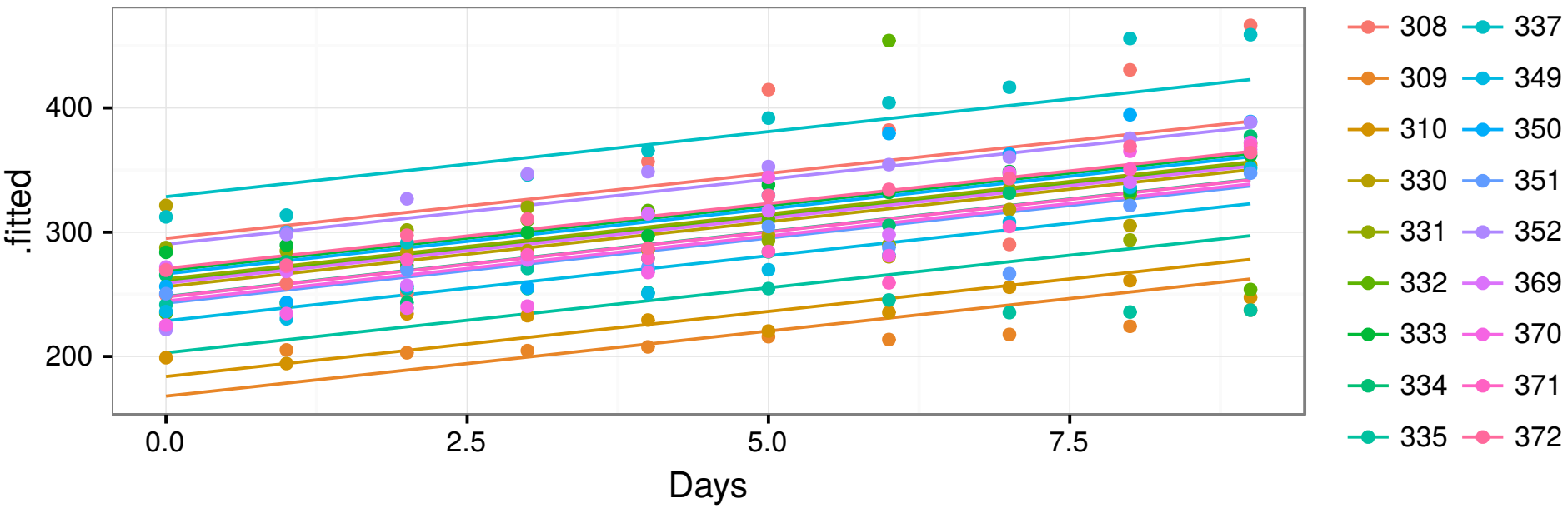
$$\mathbf{Reaction} = \mathbf{X}\beta + \varepsilon$$

```
W2 <- 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`, $\sigma$ — 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.

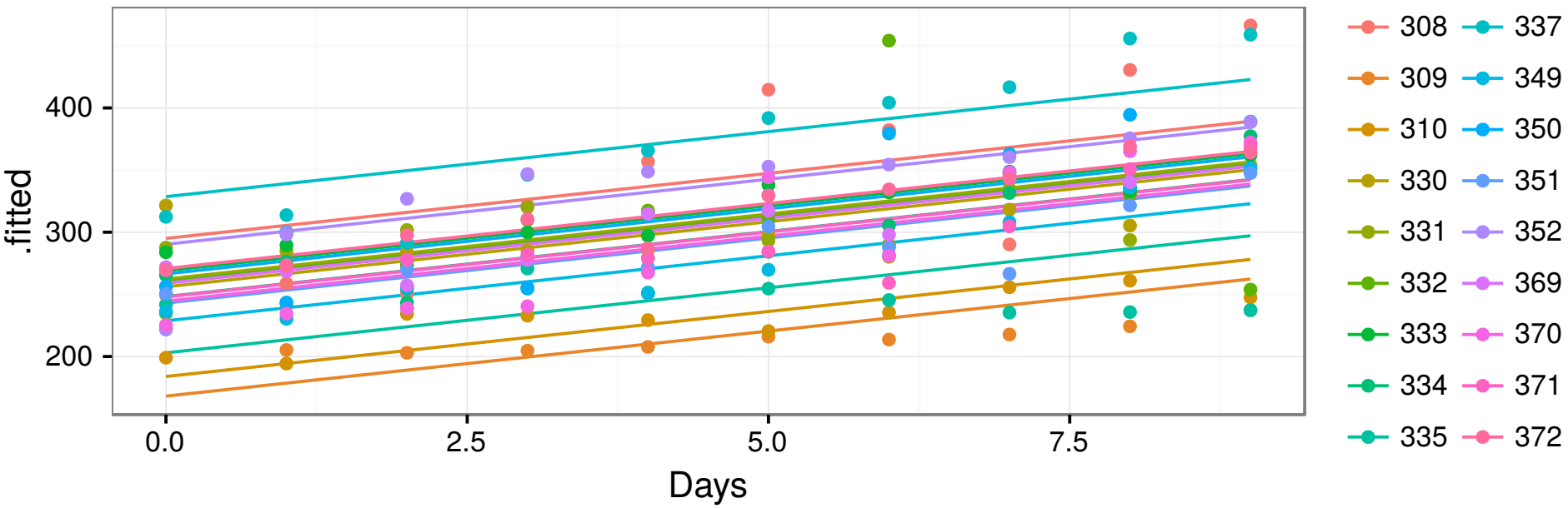
# The Ugly. Что нам делать с этим множеством прямых?

```
W2_diag <- fortify(W2)
ggplot(W2_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?

```
W2_diag <- fortify(W2)
ggplot(W2_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))
```




- Нас не интересует, как различается время реакции каждого конкретного субъекта. Можем попытаться вместо подбора отдельных интерсептов, оценить разброс их значений.

# Можно посмотреть на группирующий фактор иначе!

Нам не важны конкретные значения на разных уровнях фактора. Мы можем представить, что эффект фактора — случайная величина. Мы 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 $H_0 : \mu_1 = \mu_2 = \ldots = \mu_i = \mu$ | 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 $\sigma$ нужно 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

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

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

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

$\varepsilon_i \sim N(0, \Sigma)$ — остатки модели нормально распределены со средним 0 и матрицей ковариаций $\Sigma_i$ (дисперсией $\sigma^2$)

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

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

# В примере модель со случайным отрезком можно записать так:

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

$b_i \sim N(0, d^2)$ — случайный эффект субъекта (intercept)
$\varepsilon_{ij} \sim N(0, \sigma^2)$ — остатки модели
$i = 1, 2, ..., 18$ — субъекты
$j = 1, 2, ..., 10$ — дни

# В примере модель со случайным отрезком можно записать так:

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

$b_i \sim N(0, d^2)$ — случайный эффект субъекта (intercept)
$\varepsilon_{ij} \sim N(0, \sigma^2)$ — остатки модели
$i = 1, 2, ..., 18$ — субъекты
$j = 1, 2, ..., 10$ — 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:

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

$b_i \sim N(0, d^2)$ — случайный эффект субъекта (intercept)
$\varepsilon_{ij} \sim N(0, \sigma^2)$ — остатки модели
$i = 1, 2, ..., 18$ — субъекты
$j = 1, 2, ..., 10$ — 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:

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

# Теперь разберемся с допущениями модели

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

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

# Теперь разберемся с допущениями модели

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

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

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

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

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

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

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

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

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

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

# Матрица ковариаций переменной-отклика

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

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

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

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

Матрица ковариаций переменной-отклика:

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

$\mathbf{D}$ — 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:

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

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

# Индуцированная корреляция - следствие включения в модель случайных эффектов

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

# Индуцированная корреляция - следствие включения в модель случайных эффектов

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

$d^2$ — ковариация между наблюдениями одного субъекта
$\sigma^2 + d^2$ — дисперсия

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

# Индуцированная корреляция - следствие включения в модель случайных эффектов

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

$d^2$ — ковариация между наблюдениями одного субъекта
$\sigma^2 + d^2$ — дисперсия

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

## Коэффициент внутриклассовой корреляции $d^2/(\sigma^2 + d^2)$

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 | lme() из nlme | lmer() из lme4 | glmer() из lme4 | glmmPQL() из MASS |
|---|---|---|---|---|
| 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) | + | - | - | + |

# Синтаксис для смешанных моделей в R

**Фиксированная часть модели** задается обычной двухсторонней формулой

```
Y ~ 1 + X1 + ... + Xn
```

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

```
~ 1 + X1 + ... + Xn |A
```

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

```
~ 1 + X1 + ... + Xn |A/B/C
```

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

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

# Смешанные модели со случайным отрезком в R

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

```
detach("package:lme4") # выгружаем lme4, из которого мы взяли данные, чтобы н
library(nlme)
M1 <- lme(Reaction ~ Days, random = ~ 1 | Subject, data = sl)
```

Что дальше?

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

```
detach("package:lme4") # выгружаем lme4, из которого мы взяли данные, чтобы не
library(nlme)
M1 <- lme(Reaction ~ Days, random = ~ 1 | Subject, data = sl)
```

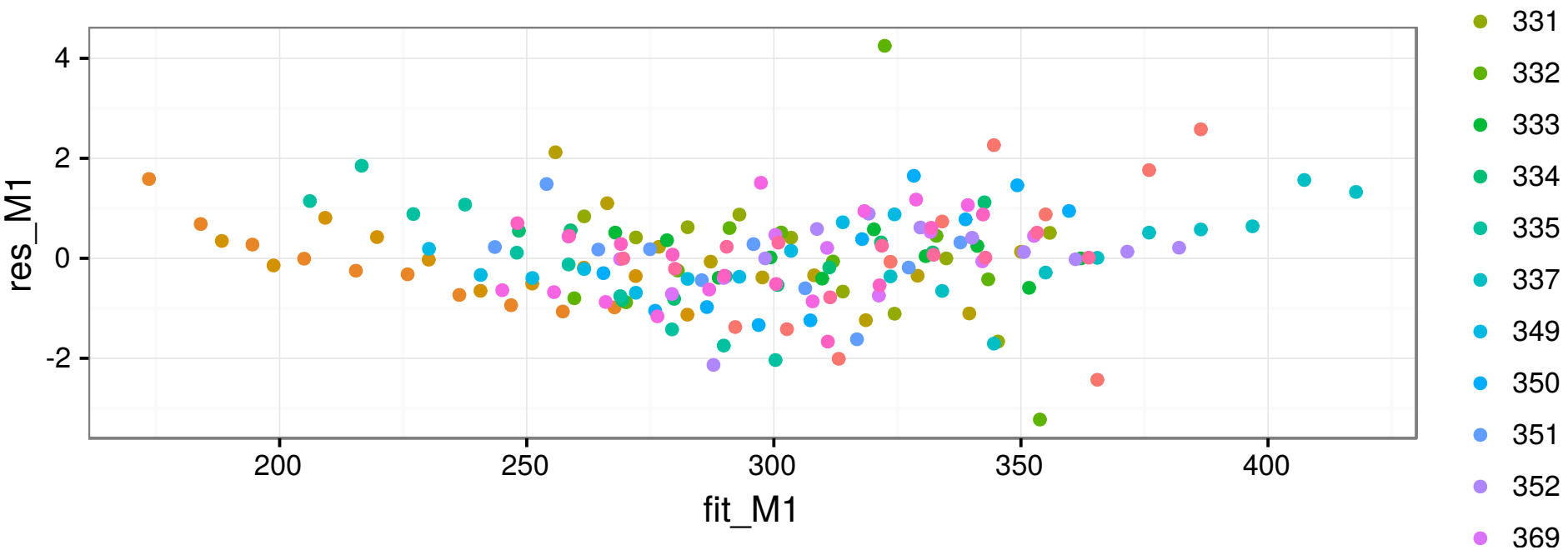
Что дальше?

Правильно, анализ остатков

# 1. Анализ остатков

1. График остатков от предсказанных значений

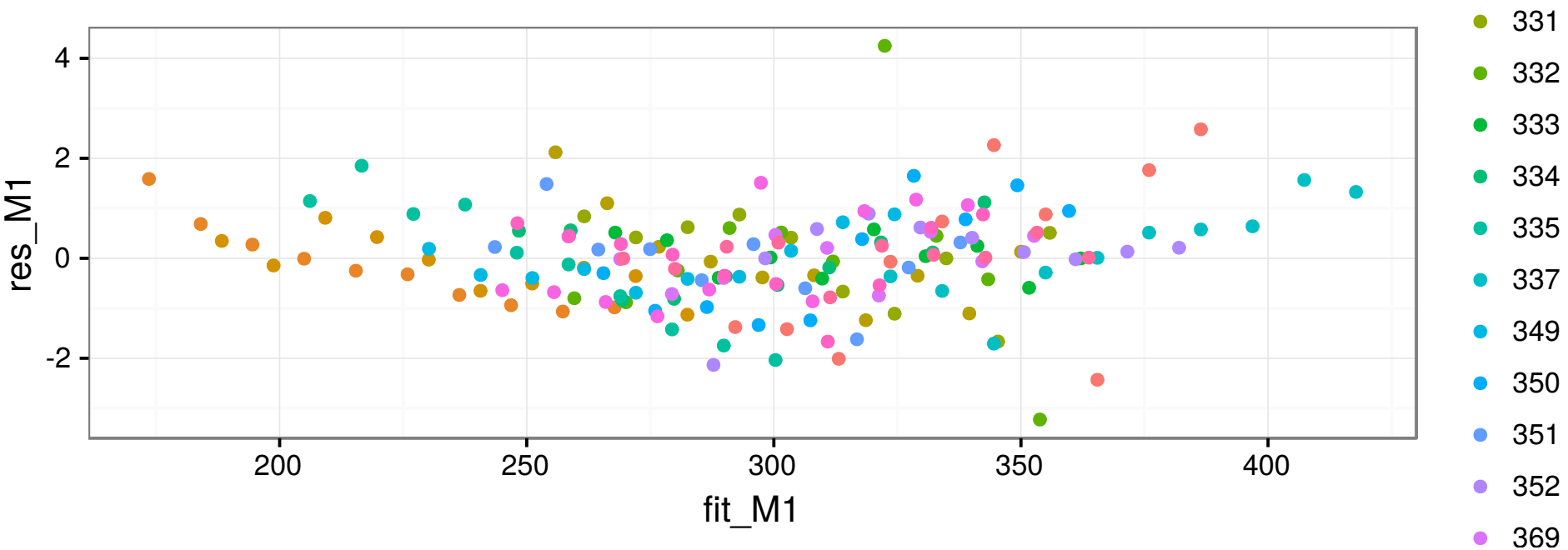
```
# plot(M1)
sl$res_M1 <- resid(M1, type = "pearson")
sl$fit_M1 <- fitted(M1)
ggplot(sl) + geom_point(aes(x = fit_M1, y = res_M1, colour = Subject))
```



# 1. Анализ остатков

1. График остатков от предсказанных значений

```
# plot(M1)
sl$res_M1 <- resid(M1, type = "pearson")
sl$fit_M1 <- fitted(M1)
ggplot(sl) + geom_point(aes(x = fit_M1, y = res_M1, colour = Subject))
```

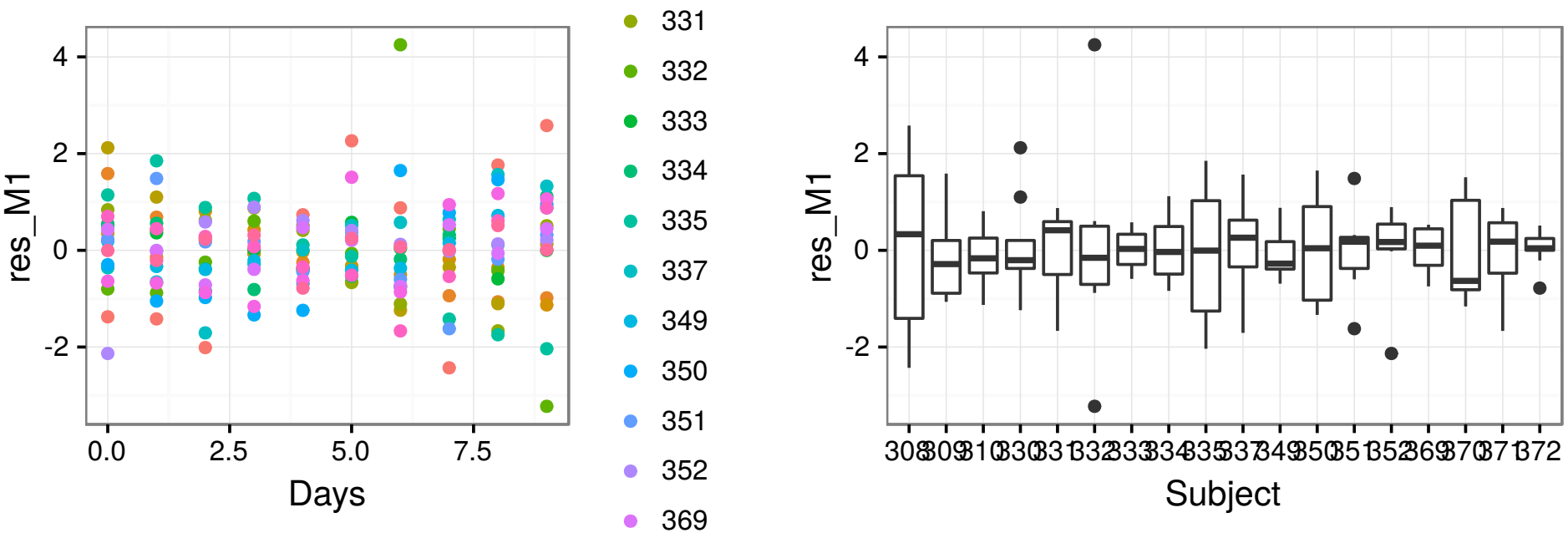



- Есть большие остатки, гетерогенность дисперсий

# 1. Анализ остатков

2. График остатков от ковариат в модели

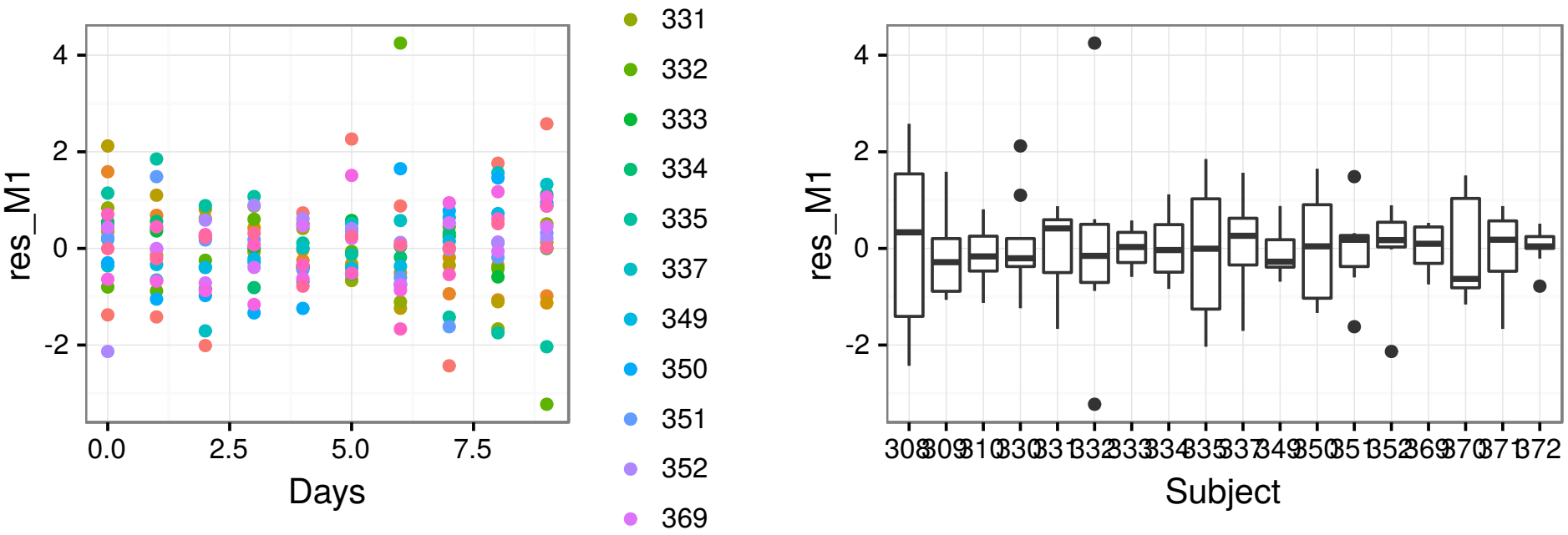
```
p <- ggplot(data = sl, aes(y = res_M1))
library(gridExtra)
grid.arrange(p + geom_point(aes(x = Days, colour = Subject)),
             p + geom_boxplot(aes(x = Subject)),
             ncol = 2)
```



# 1. Анализ остатков

2. График остатков от ковариат в модели

```
p <- ggplot(data = sl, aes(y = res_M1))
library(gridExtra)
grid.arrange(p + geom_point(aes(x = Days, colour = Subject)),
             p + geom_boxplot(aes(x = Subject)),
             ncol = 2)
```




- Большие остатки у наблюдений для 332 субъекта
- Гетерогенность дисперсий
- Пока оставим все как есть — на следующем занятии мы научимся моделировать гетерогенность дисперсии.

# 2. Проверка влияния факторов

Достаточно **одного** из этих трех вариантов.

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

(а) По значениям t-(или -z) статистики (по REML оценке)

(б) F-критерий - приблизительный результат (REML оценка)

(в) likelihood ratio test или AIC (ML оценка)

- Либо попарное сравнение вложенных моделей при помощи likelihood ratio test
- Либо сравнение моделей по AIC

# 2(a) По значениям t-(или -z) статистики (по REML оценке)

Подходит для непрерывных переменных или факторов с 2 уровнями.
Дает приблизительный результат.

```
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
#
```

# 2(б) F-критерий - приблизительный результат (REML оценка)

Осторожно с интерпретацией!

- `anova()` — Type I SS
- `Anova()` из пакета `car` — Type II, III SS

```
anova(M1)
```

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

# 2(б) F-критерий - приблизительный результат (REML оценка)

Осторожно с интерпретацией!

- `anova()` — Type I SS
- `Anova()` из пакета `car` — Type II, III SS

```
anova(M1)
```

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

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

# 2(в1) Попарное сравнение вложенных моделей при помощи likelihood ratio test

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 <- update(M1.ml, . ~ . - Days)
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
```

df теста - это разница df сравниваемых моделей = 4 - 3 = 1

# 2(в1) Попарное сравнение вложенных моделей при помощи likelihood ratio test

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 <- update(M1.ml, . ~ . - Days)
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
```

df теста - это разница df сравниваемых моделей = 4 - 3 = 1

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

# 2(в2) Сравнение моделей по AIC

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

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

# 2(в2) Сравнение моделей по AIC

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

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

- Продолжительность бессонницы влияет на время реакции (AIC)

# 3. Представление результатов

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

REML оценка параметров более точна (оценка случайных факторов)

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

Для проверки финальной модели необходимо провести анализ остатков (те же графики, что и в п.1). Поскольку модель не изменилась, не привожу их здесь

# Вычисляем внутриклассовую корреляцию

$$sigma^2_{Subject} / (sigma^2_{Subject} + sigma^2)$$

```
M1_fin
```

```
В результатах
Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    37.12383 30.99123
```

```
# Внутриклассовая корреляция
37.12383^2 / (37.12383^2 + 30.99123^2)
```

```
# [1] 0.589309
```

# Вычисляем внутриклассовую корреляцию

$$sigma^2_{Subject} / (sigma^2_{Subject} + sigma^2)$$

```
M1_fin
```

```
В результатах
Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    37.12383 30.99123
```

```
# Внутриклассовая корреляция
37.12383^2 / (37.12383^2 + 30.99123^2)
```

```
# [1] 0.589309
```

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

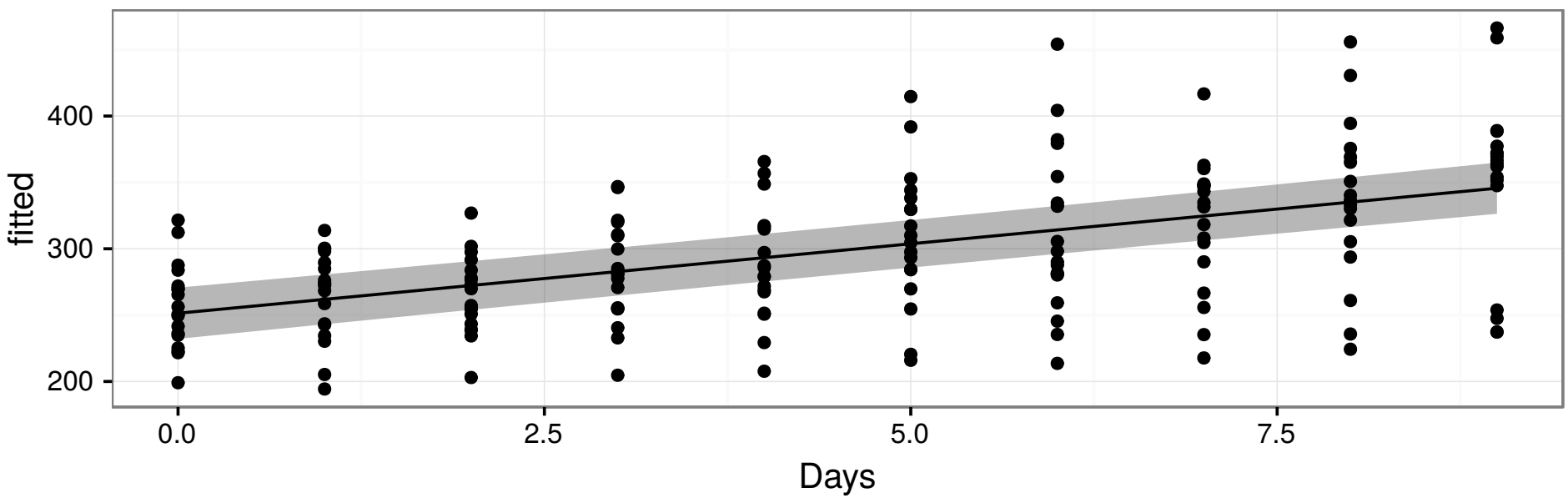
# График предсказанных значений для результатов

1-й вариант — предсказания по фиксированной части модели

```
library(plyr)
MyData_M1 <- ddply(
  sl, .(Subject), summarise,
  Days = seq(min(Days), max(Days), length = 10)
  )
# level = 0 - для фиксированных эффектов (т.е. без учета субъекта)
MyData_M1$fitted <- predict(M1_fin, MyData_M1, level = 0)

# или то же самое при помощи матриц
X <- model.matrix(~ Days, data = MyData_M1)
betas <- fixef(M1_fin)
MyData_M1$fitted <- X %*% betas

# стандартные ошибки и дов. интервалы
MyData_M1$se <- sqrt( diag(X %*% vcov(M1_fin) %*% t(X)) )
MyData_M1$lwr <- MyData_M1$fitted - 1.98 * MyData_M1$se
MyData_M1$upr <- MyData_M1$fitted + 1.98 * MyData_M1$se
```

# 1-й вариант. График с предсказаниями по фиксированной части модели

```
ggplot(data = MyData_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))
```



# График предсказанных значений для результатов

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

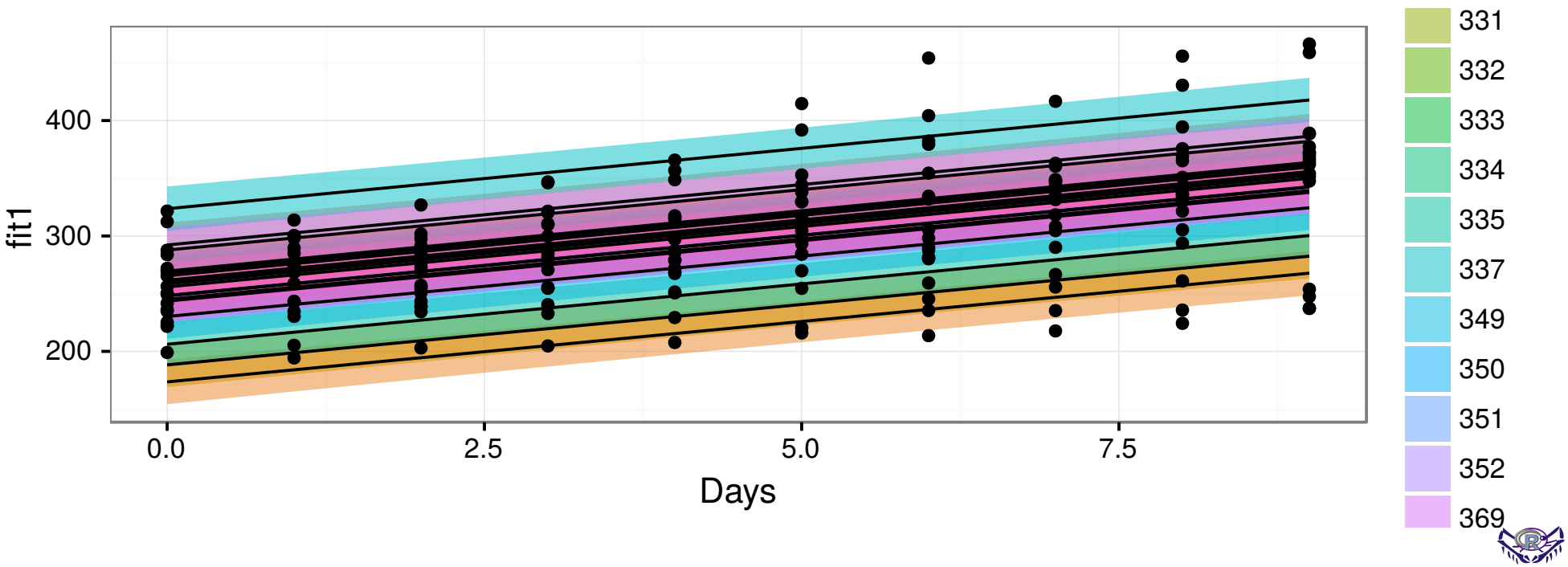
2-й вариант — предсказания для каждого субъекта

beta_0 + beta * Days + случайный эффект субъекта

```
MyData_M1$fit1 <- predict(M1_fin, MyData_M1, level = 1)
# или то же самое при помощи матриц
# случайные эффекты для каждого субъекта
# это датафрейм с одним столбцом
rand <- ranef(M1_fin)
# "разворачиваем" для каждой строки данных
all_rand <- rand[as.numeric(MyData_M1$Subject), 1]
# прибавляем случайные эффекты к предсказаниям фикс. части
MyData_M1$fit1 <- X %*% betas + all_rand
```

# 2-й вариант. График с предсказаниями для индивидуальных уровней случайного фактора

```
ggplot(MyData_M1, aes(x = Days, y = fit1, group = Subject)) +
  geom_ribbon(alpha = 0.5, aes(fill = Subject, ymin = fit1 - 1.98*se,
                  ymax = fit1 + 1.98*se)) +
  geom_line() +
  geom_point(data = sl, aes(x = Days, y = Reaction))
  # попробуйте добавить facet_wrap(~Subject)
```



# 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.

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

Дальнейшие действия по прежнему плану:

- Анализ остатков
- Проверка влияния факторов + подбор оптимальной модели
- Анализ остатков финальной модели
- Визуализация предсказаний

# Задание

Проверьте получившуюся модель MS1

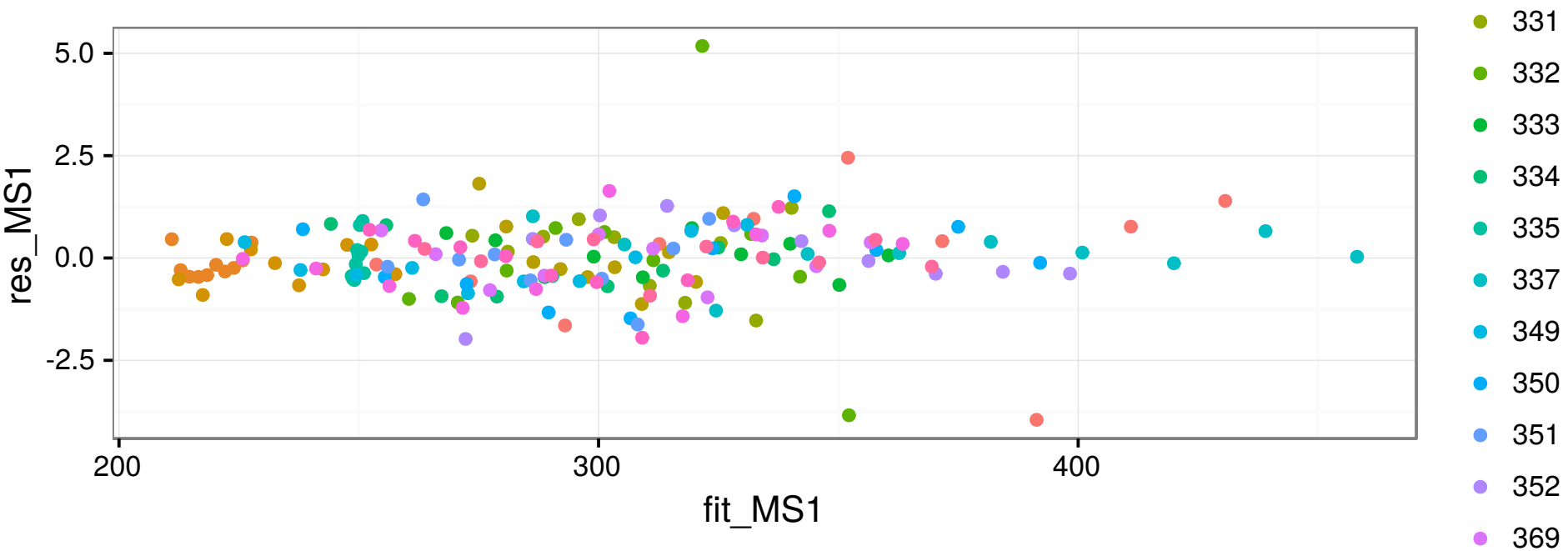
Сделайте самостоятельно:

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

# Решение: 1. Анализ остатков

1. График остатков от предсказанных значений

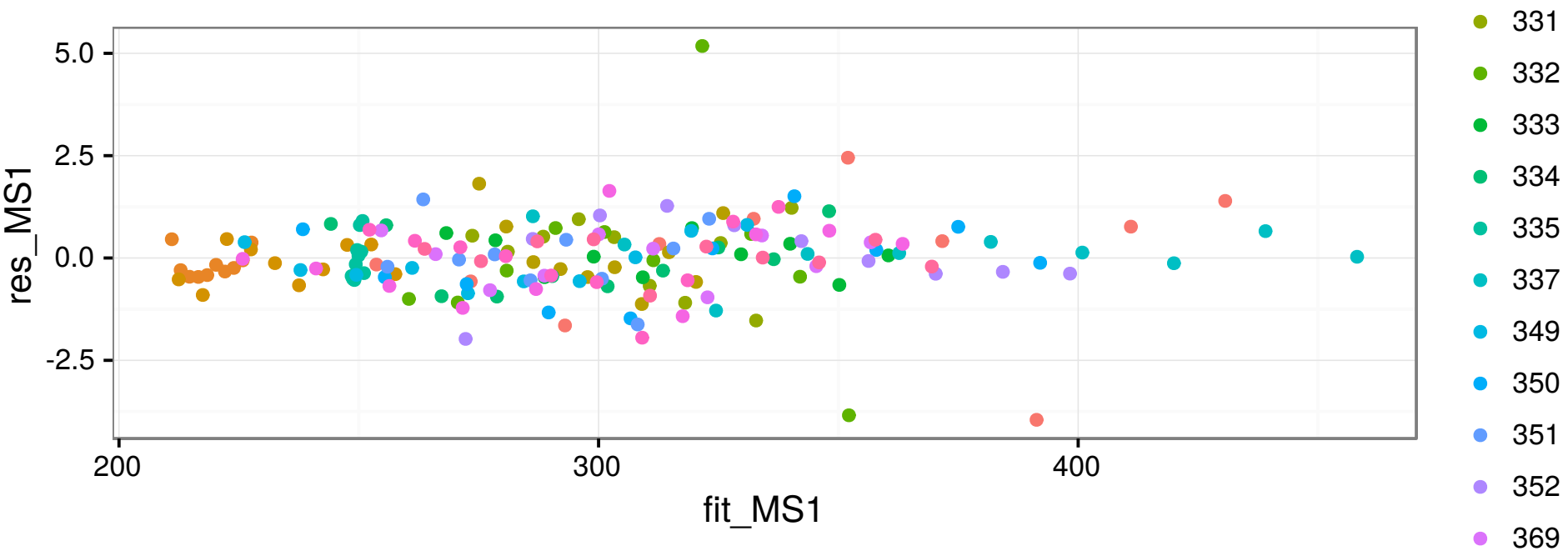
```
# plot(M1)
sl$res_MS1 <- resid(MS1, type = "pearson")
sl$fit_MS1 <- fitted(MS1)
ggplot(sl) + geom_point(aes(x = fit_MS1, y = res_MS1, colour = Subject))
```



# Решение: 1. Анализ остатков

1. График остатков от предсказанных значений

```
# plot(M1)
sl$res_MS1 <- resid(MS1, type = "pearson")
sl$fit_MS1 <- fitted(MS1)
ggplot(sl) + geom_point(aes(x = fit_MS1, y = res_MS1, colour = Subject))
```

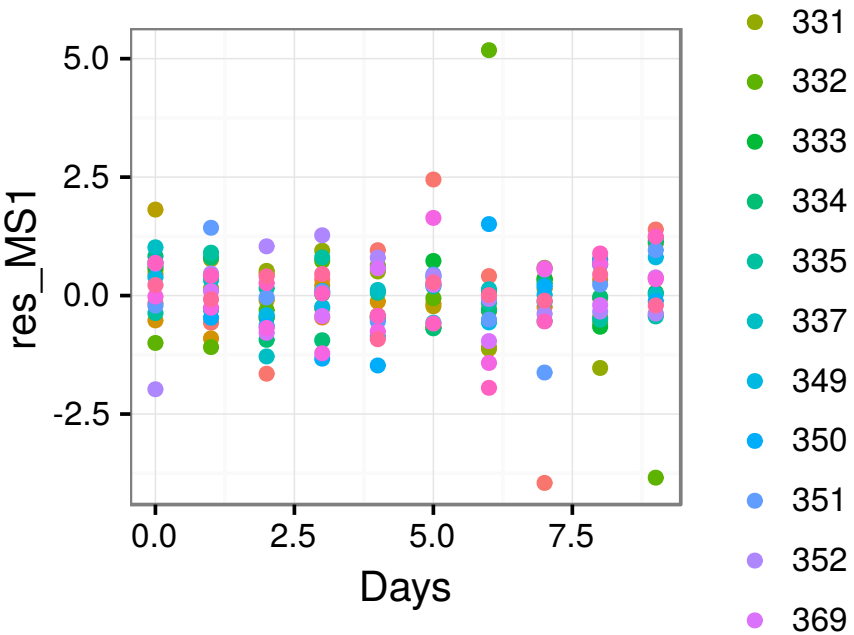


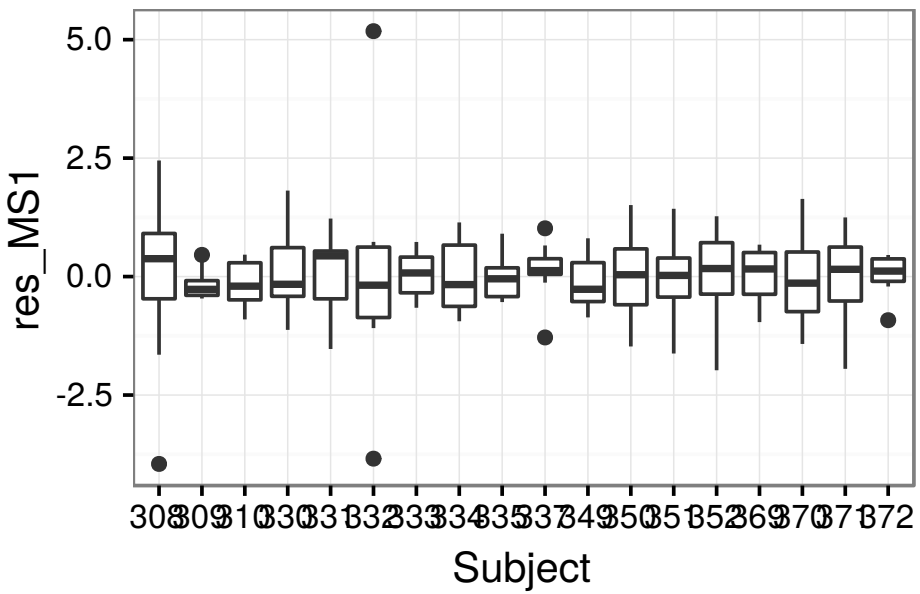

- Есть большие остатки, гетерогенность дисперсий не выражена. Стало явно лучше, чем было.

# Решение: 1. Анализ остатков

2. График остатков от ковариат в модели

```
p <- ggplot(data = sl, aes(y = res_MS1))
grid.arrange(p + geom_point(aes(x = Days, colour = Subject)),
             p + geom_boxplot(aes(x = Subject)),
             ncol = 2)
```

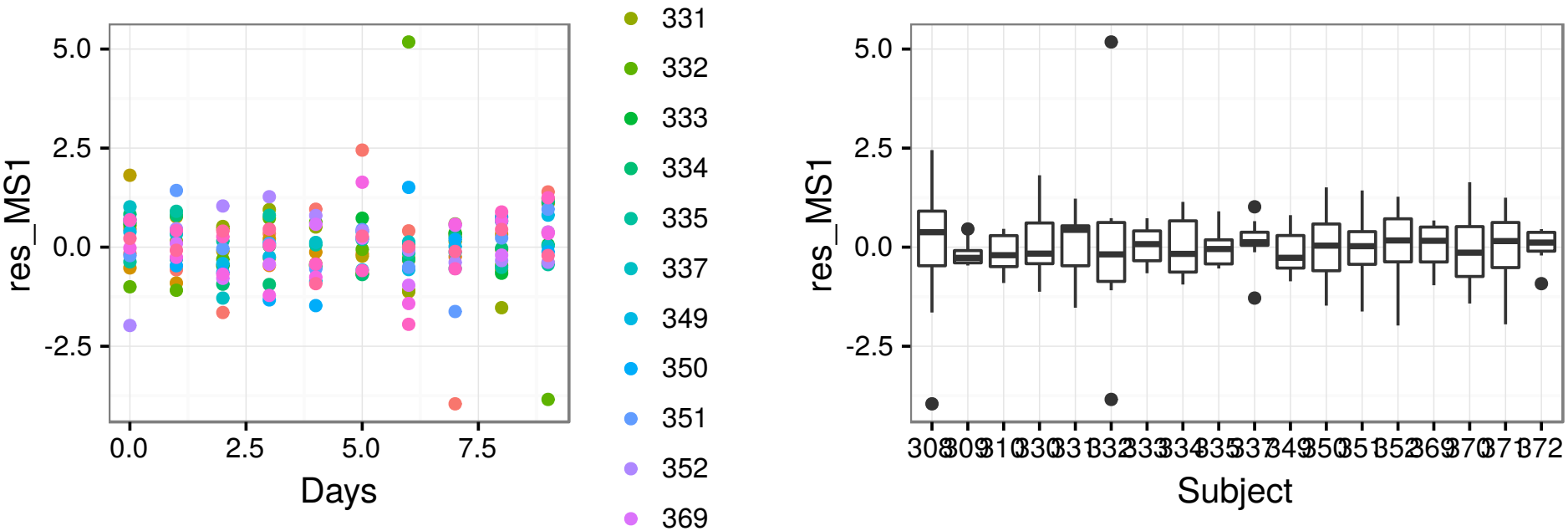


# Решение: 1. Анализ остатков

2. График остатков от ковариат в модели

```
p <- ggplot(data = sl, aes(y = res_MS1))
grid.arrange(p + geom_point(aes(x = Days, colour = Subject)),
             p + geom_boxplot(aes(x = Subject)),
             ncol = 2)
```




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

# Решение: 2. Проверка влияния факторов (Days)

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: 2. Проверка влияния факторов (Days)

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 = 23, df = 1, p < 0.01).

# Решение: 2. Проверка влияния факторов (Days)

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 = 23, df = 1, $p < 0.01$).

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: 2. 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: 2. 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 = 42, df = 2, p < 0.01)

# Решение: 3. 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,
               method = "REML", data = sl)
```

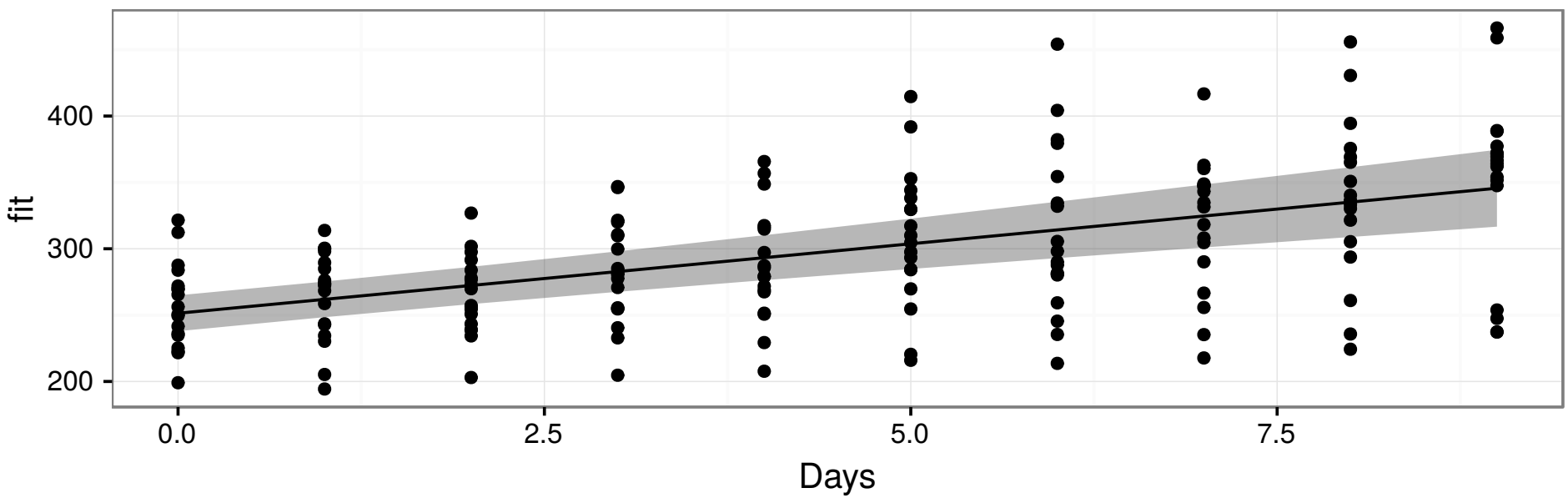
# Решение: График предсказанных значений для результатов

1-й вариант — предсказания по фиксированной части модели

```
MyData_MS1 <- ddply(
  sl, .(Subject), summarise,
  Days = seq(min(Days), max(Days), length = 10)
  )
# level = 0 - для фиксированных эффектов (т.е. без учета субъекта)
MyData_MS1$fit <- predict(MS1_fin, MyData_MS1, level = 0)

# или то же самое при помощи матриц
X <- model.matrix(~ Days, data = MyData_MS1)
betas = fixef(MS1_fin)
MyData_MS1$fit <- X %*% betas

# стандартные ошибки и дов. интервалы
MyData_MS1$se <- sqrt( diag(X %*% vcov(MS1_fin) %*% t(X)) )
MyData_MS1$lwr <- MyData_MS1$fit - 1.98 * MyData_MS1$se
MyData_MS1$upr <- MyData_MS1$fit + 1.98 * MyData_MS1$se
```

# Решение: 1-й вариант. График с предсказаниями по фиксированной части модели

```
ggplot(data = MyData_MS1, 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))
```



# Решение: График предсказанных значений для результатов

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

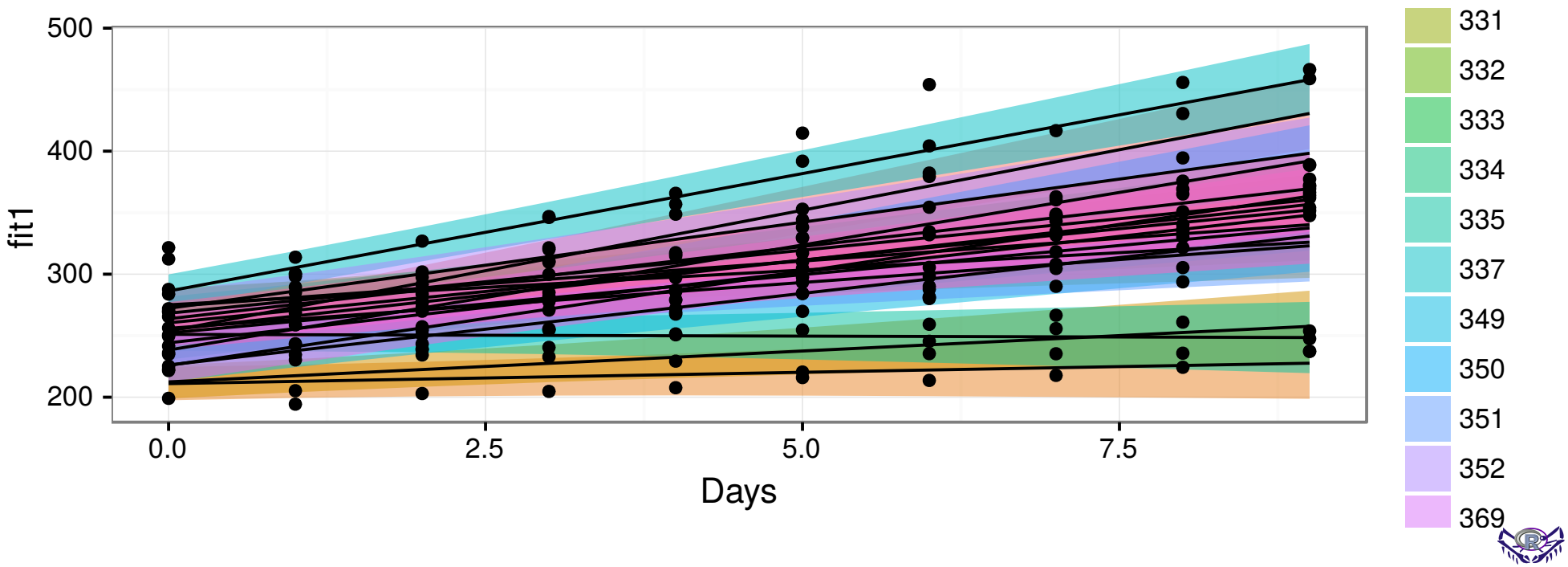
2-й вариант — предсказания для каждого субъекта

beta_0 + beta * Days + случайный эффект субъекта

```
MyData_MS1$fit1 <- predict(MS1_fin, MyData_MS1, level = 1)
# или то же самое при помощи матриц
# случайные эффекты для каждого субъекта
# это датафрейм с двумя столбцами
rand <- ranef(MS1_fin)
# "разворачиваем" для каждой строки данных
all_rand <- rand[as.numeric(MyData_MS1$Subject), ]
# прибавляем случайные эффекты к предсказаниям фикс. части
MyData_MS1$fit1 <- (betas[1] + all_rand[, 1]) + (betas[2] + all_rand[, 2]) * I
```

# Решение: 2-й вариант. График с предсказаниями для индивидуальных уровней случайного фактора

```
ggplot(MyData_MS1, aes(x = Days, y = fit1, group = Subject)) +
  geom_ribbon(alpha = 0.5, aes(fill = Subject, ymin = fit1 - 1.98*se,
                  ymax = fit1 + 1.98*se)) +
  geom_line() +
  geom_point(data = sl, aes(x = Days, y = Reaction))
  # попробуйте добавить facet_wrap(~Subject)
```



# 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.
- Коэффициент внутриклассовой корреляции оценивает, насколько коррелируют друг с другом наблюдения из одной и той же группы случайного фактора.

# Дополнительные ресурсы


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