Смешанные линейные модели для бинарных данных

Линейные модели…

Вадим Хайтов, Марина Варфоломеева

Что мы знаем про моделирование бинарных данных

Бинарные данные вокруг нас

Мы встречаемся с бинарными исходами очень часто

- Проголосовали за данного кандидата или не проголосовали
- Прогноз сбылся или не сбылся
- Человек нравится или не нравится
- Состояние пациента улучшилось или ухудшилось

Бинарные данные – очень распространенный тип зависимых переменных, они отражают соотношение положительных и отрицательных исходов в ответ на влияние предикторов.

Бинарные данные тоже могут быть связаны со случайными эффектами

Бинарные данные могут находиться в связи с группирующими, случайными факторами.

Например, нас интересует как зависит голосование за или против кандидата от пола, возраста и материального положения избирателя.

Задание: Придумайте, какие могут быть группирующие (случайные) факторы в подобном исслдеовании.

Бинарные данные тоже могут быть связаны со случайными эффектами

Бинарные данные могут находиться в связи с группирующими, случайными факторами.

Например, нас интересует как зависит голосование за или против кандидата от пола, возраста и материального положения избирателя.

Задание: Придумайте, какие могут быть группирующие (случайные) факторы в подобном исслдеовании.

В качестве группирующих факторов, не входящих в интерес исследователя, могут выступать регион, город, предприятие, к которым относятся избиратели. Все это заставляет нас рассмотреть обобщенные смешанные линейные модели (GLMM).

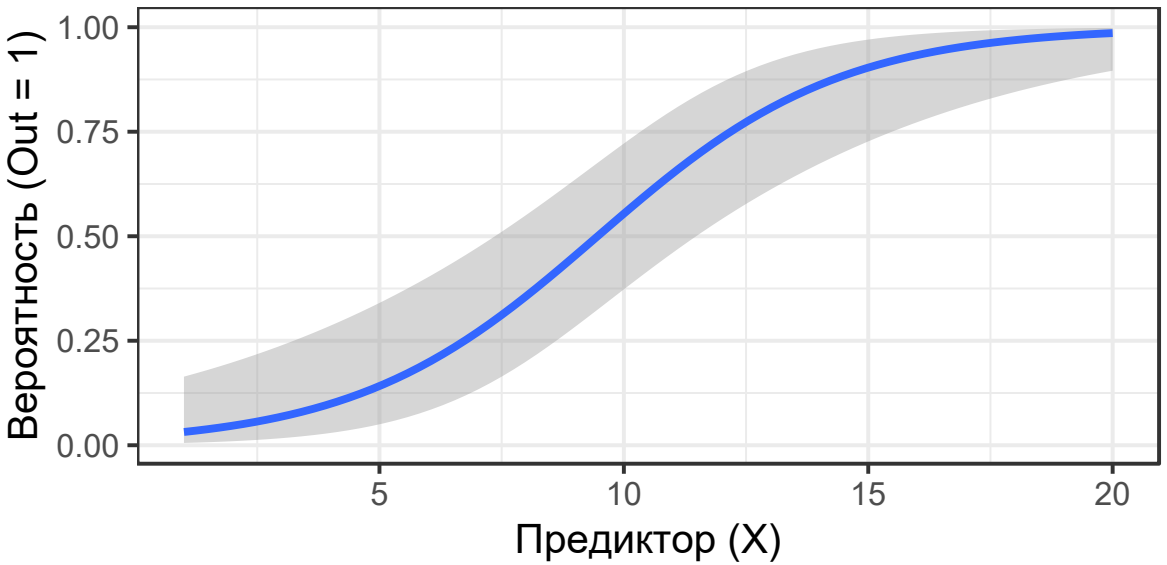
Вспомним, как устроена работа с линейными моделями, описывающими поведение бинарных данных.

Структура данных для анализа

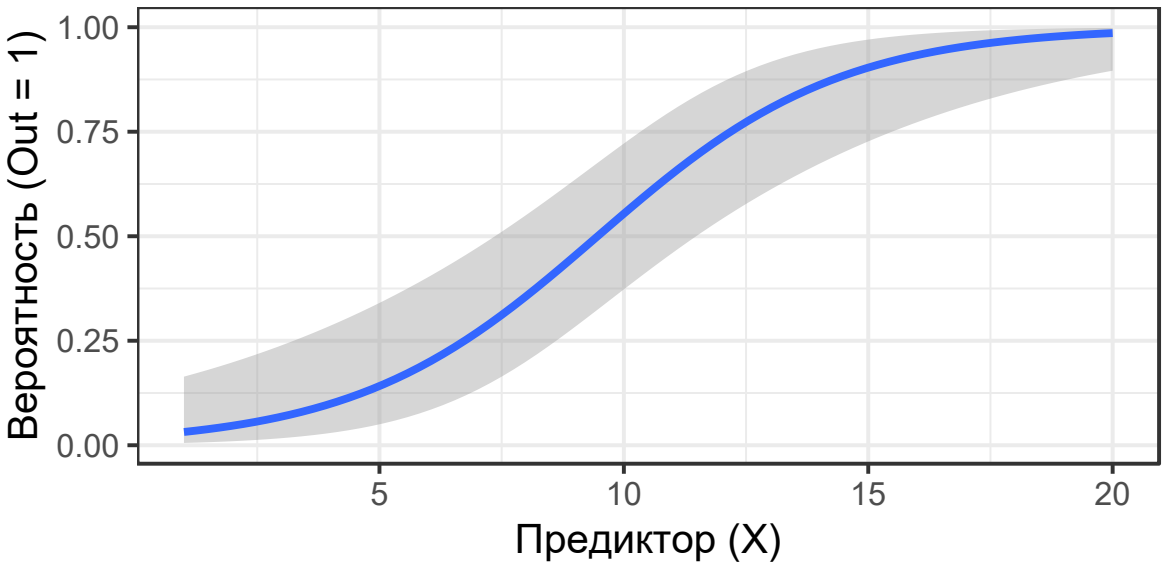
- При разных значениях предиктора, зависимая переменная принимает значения 1 или 0 (событие произошло или не произошло)
- Предиктор может быть непрерывным или дискретным.
- Предикторов может быть один или много.

```
#    X Out
# 1  1   0
# 2  1   0
# 3  1   0
# 4  2   0
# 5  2   0
# 6  2   0
# 7  3   0
# 8  3   0
# 9  3   1
# 10 4   0
# 11 4   0
# 12 4   0
# 13 5   0
# 14 5   0
# 15 5   1
```

Цель анализа - логистическая кривая



Цель анализа - логистическая кривая



- Вместо перменной отклика в виде 1 или 0 переходим к вероятностям событий (π), распределенным от 0 до 1.
- Итогом анализа является логистическая регрессионная кривая, связывающая значения предиктора и вероятность появления события $Out_i = 1$.

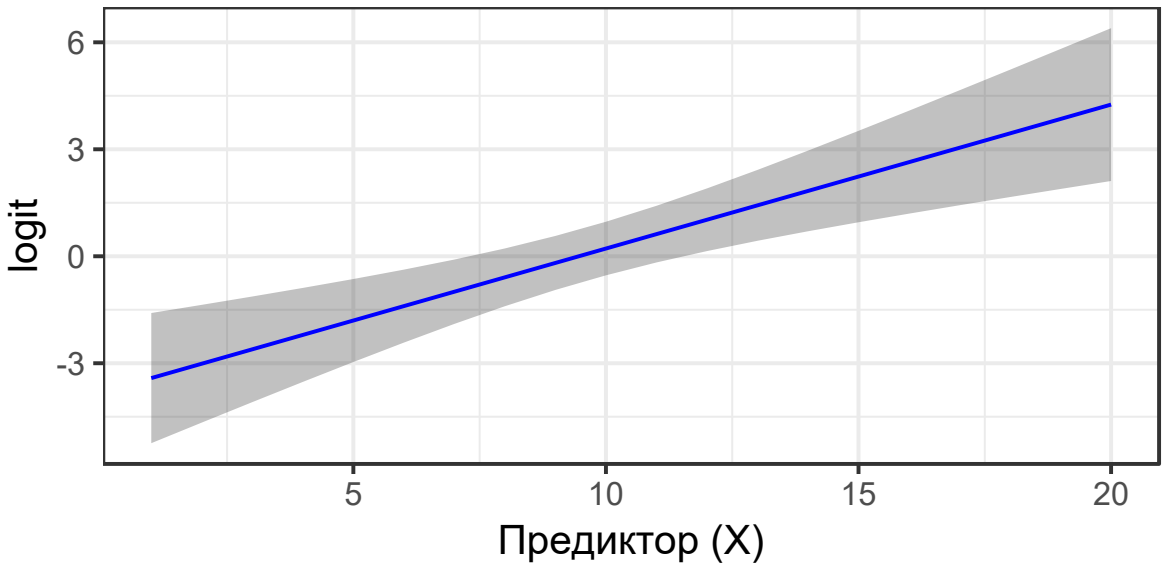
Вероятности, шансы, логиты

- Для подбора параметров логистической кривой необходимо перейти к линейной зависимости.

Вероятности, шансы, логиты

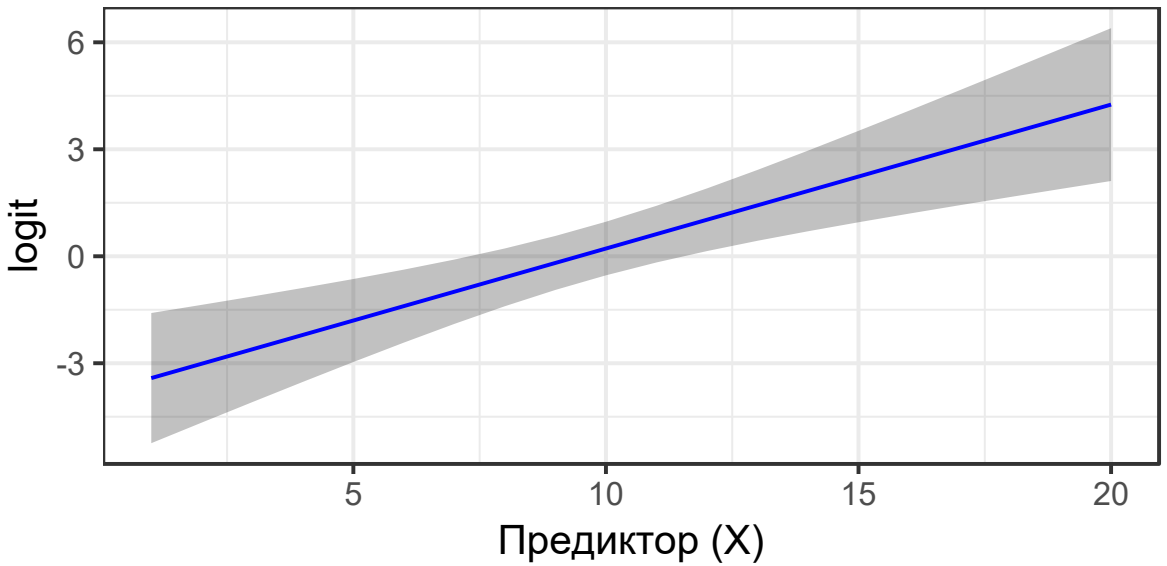
- Для подбора параметров логистической кривой необходимо перейти к линейной зависимости.
- Вместо вероятностей, переходим к отношению шансов: $odds = \frac{\pi}{1-\pi}$, распределенным от 0 до $+\infty$.
- Далее вместо отношения шансов используем логиты: $ln(odds) = ln(\frac{\pi}{1-\pi})$, которые варьируют от $-\infty$ до $+\infty$.

Связь лгитов с предиктором



- Логит-связывающая функция:

Связь лгитов с предиктором



- Логит-связывающая функция: логиты линейно связаны с предиктором:

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Обобщенная лиенйная модель для бинарного отклика

$$y_i \sim Binomial(n = 1, \pi_i)$$

$$E(y_i) = \pi_i = \frac{e^{\beta_0 + \beta_1 x_1 + ... + \beta_{p-1}\ x_{p-1}}}{1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_{p-1}\ x_{p-1}}}$$

$ln(\frac{\pi_i}{1-\pi_i}) = \eta_i$ — функция связи логит, переводит вероятности в логиты.

$$\eta_i = \beta_0 + \beta_1 x_{1\,i} + ... + \beta_{p-1\,i}\ x_{p-1\,i}$$

- Для подбора параметров регрессионной модели с бинарным откликом используют обобщенные линейные модели (GLM или GLMM).
- Подбор параметров модели основан на методе максимального правдоподобия.

```
model <-  glm(Out ~ X, data = dat, family = binomial(link="logit") )
```

Для диагностики модели с бинарным откликом необходимо проверить

1. Независимость испытаний друг от друга (определяется дизайном сбора материала).

Для диагностики модели с бинарным откликом необходимо проверить

1. Независимость испытаний друг от друга (определяется дизайном сбора материала).
2. Линейность связи с предиктором (проверяется по графику рассеяния остатков).
3. Если есть предикторы, которые не вошли в модель, то наличие паттерна в остатках по отношению к ним может говорить о нарушении линейности связи или нарушении независимости испытаний (часто выявляется при анализе данных, связанных с временными рядами).

Для диагностики модели с бинарным откликом необходимо проверить

1. Независимость испытаний друг от друга (определяется дизайном сбора материала).
2. Линейность связи с предиктором (проверяется по графику рассеяния остатков).
3. Если есть предикторы, которые не вошли в модель, то наличие паттерна в остатках по отношению к ним может говорить о нарушении линейности связи или нарушении независимости испытаний (часто выявляется при анализе данных, связанных с временными рядами).
4. Отсутствие избыточности дисперсии (проверяется по соотношению суммы квадратов пирсоновских остатков и числа степеней свободы).

Смысл коэффициентов модели

```
summary(model)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.82050    1.00162  -3.814 0.000137 ***
X            0.40372    0.09666   4.177 2.96e-05 ***
```

Смысл коэффициентов модели

```
summary(model)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.82050    1.00162  -3.814 0.000137 ***
X            0.40372    0.09666   4.177 2.96e-05 ***
```

Поскольку коэффициент b_1 положительный, то при увеличении предиктора на единицу отношение шансов возрастает в e^{b_1} = exp(0.4037184) = 1.5 раз.

Модель

Подобранные параметры

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.82050    1.00162  -3.814 0.000137 ***
X            0.40372    0.09666   4.177 2.96e-05 ***
```

Связь, переменной отклика с предиктором описывается следующей моделью

$$E(y_i) = \pi_i = \frac{e^{-3.8+0.4X_i}}{1 + e^{-3.8+0.4X_i}}$$

$$y_i \sim Binomial(n = 1, \pi_i)$$

Пример – морские звезды и мидии

Различают ли морские звезды два вида мидий: предыстория

В Белое море, где обитают ценные промысловые моллюски, атлантические мидии, недавно вселились мидии тихоокеанские. Важно понять, могут ли хищные морские звезды регулировать численность вида-вселенца.

Рассматривая этот пример на предыдущей лекции, посвященной бинарным данным, мы обсуждали только данные первого, разведочного эксперимента. Было показано, что шансы быть съеденными у вида-вселенца выше, чем у коренного вида.

Данные: Khaitov et al, 2018

Случайные факторы, как инструмент поверки результатов на прочность

Один из способов добавить вес полученным выводам - это устроить проверку гипотезы на материале, включающем случайные факторы:

- Многократно повторить однотипный эксперимент.
- Желательно, чтобы эксперименты проводили несколько разных команд экспериментаторов.
- В каждом эксперименте установить несколько независимых модулей, в каждом из которых идет изучаемый процесс.

Если и в таком разнородном материале будет выявляться поведение системы, соответствующее вашей гипотезе, то значит это **закономерность**, а не игра случая.

Случайный фактор первого уровня

Предыстория: разведочный эксперимент проводили в четырех независимых контейнерах. Мы включили фактор Box в фиксированную часть модели. Однако влияние этого фактора вне интереса исследования.

По своей природе фактор Box — случайный. Для надежности оценки случайного эффекта лучше использовать большое количество градаций этого фактора.

Для проведения нового исследования мы использовали от 9 до 12 экспериментальных контейнеров.

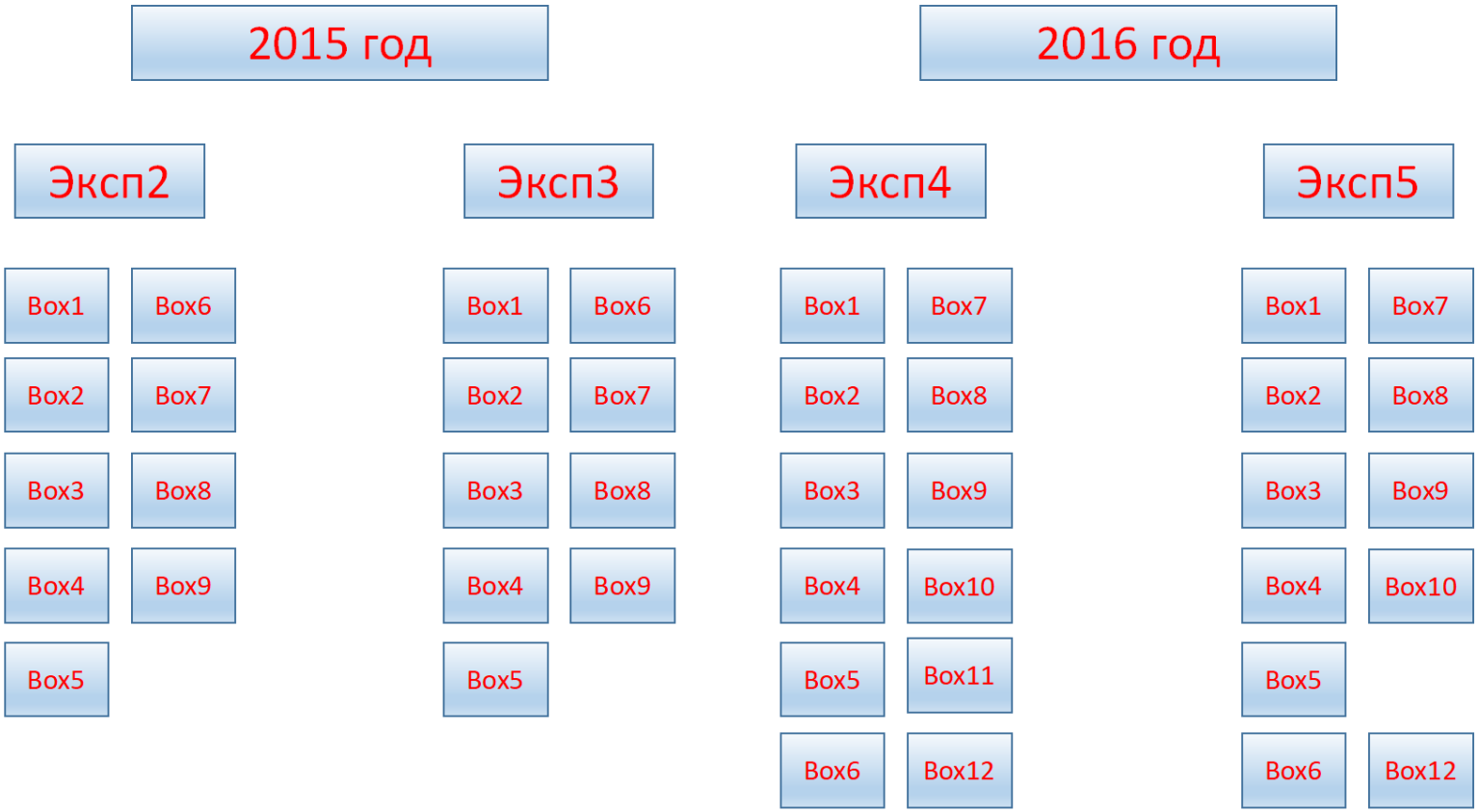
Полезно сомневаться: случайный фактор второго уровня

А не являются ли результаты одного конкретного эксперимента (даже если он однотипно воспроизводится в разных контейнерах) просто результатом счастливого стечения обстоятельств? Например, погода в период проведения одного эксперимента была какая-то особенная.

Важно повторить эксперимент несколько раз. В новом исследовании мы провели 4 эксперимента (по два 2015 и 2016 гг.).

Фактор “Эксперимент” вне интереса исследования, по своей природе — это тоже случайный фактор, но более высокого уровня, чем фактор Box.

Иерархия факторов



Случайный фактор `Box` иерархически подчинен (nested in) фактору `Experiment`.

Фактор “Год”

Фактор “Год”

Эксперимент проводили в два разных года. Для проверки нашей гипотезы важно понять устойчив ли эффект, который наблюдали в разведочном эксперименте, от года к году. Фактор `Year` надо включить в анализ, но в качестве фиксированного фактора.

Читаем данные

```
astr2 <- read.csv('data/aster_mussel_full.csv', header = TRUE)
head(astr2)
```

```
#   Year Experiment Box    L Sp   Outcome
# 1 2015       Exp2 2_1 31.2 Ed not_eaten
# 2 2015       Exp2 2_1 36.5 Tr not_eaten
# 3 2015       Exp2 2_1 37.9 Tr not_eaten
# 4 2015       Exp2 2_1 26.2 Tr not_eaten
# 5 2015       Exp2 2_1 33.3 Tr not_eaten
# 6 2015       Exp2 2_1 20.1 Tr not_eaten
```

Наводим порядок в кодировке переменных

Год закодирован числами, его надо сделать фактором

```
astr2$Year <- factor(astr2$Year)
```

Остальные категориальные переменные тоже должны быть факторами

```
astr2$Box <- factor(astr2$Box)
astr2$Sp <- factor(astr2$Sp)
```

Переменная отклика должна кодироваться как 1 (`eaten`) или 0 (`not_eaten`).

```
astr2$Out <- ifelse(test = astr2$Outcome == 'eaten', yes = 1,  no = 0)
```

Знакомимся с данными

Нет ли пропущенных значений?

```
colSums(is.na(astr2))
```

```
#       Year Experiment        Box          L         Sp    Outcome        Out 
#          0          0          0          0          0          0          0 
```

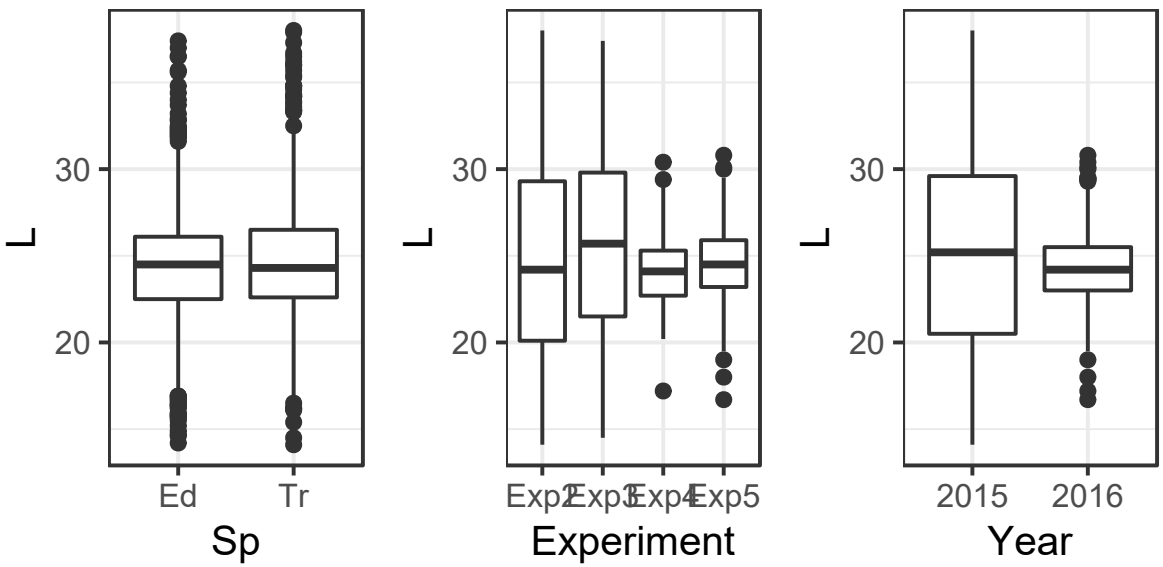
Каковы объемы выборок?

```
table(astr2$Box)
```

```
#
# 2_1 2_2 2_3 2_4 2_5 2_6 2_7 2_8 2_9 3_1 3_2 3_3 3_4 3_5 3_6 3_7 3_8 3_
#  24  21  24  23  25  22  22  24  23  24  25  23  24  22  23  24  24  24
# 4_1 4_10 4_11 4_12 4_2 4_3 4_4 4_5 4_6 4_7 4_8 4_9 5_1 5_10 5_12 5_2 5_3
#  30  30  31  30  30  30  31  29  30  30  29  29  29  36  29  30  25  36
# 5_5 5_6 5_7 5_8 5_9
#  30  29  26  28  30
```

Нет ли коллинеарности

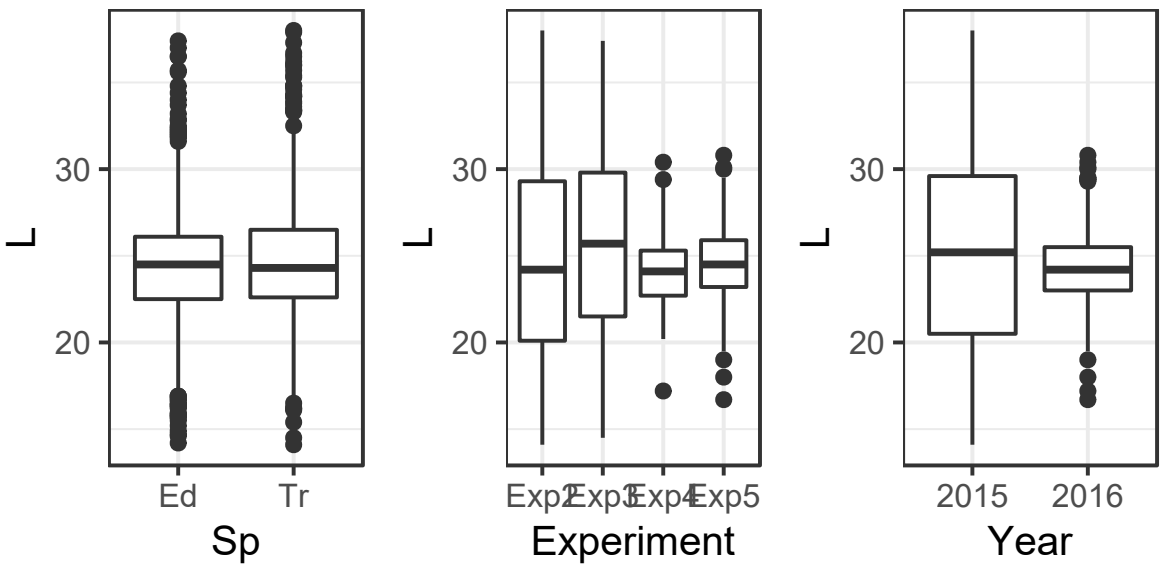
```
library(cowplot); library(ggplot2); theme_set(theme_bw())

Pl_Sp <- ggplot(astr2, aes(x = Sp, y = L)) + geom_boxplot()
Pl_exp <- ggplot(astr2, aes(x = Experiment, y = L)) + geom_boxplot()
Pl_year <- ggplot(astr2, aes(x = Year, y = L)) + geom_boxplot()
plot_grid(Pl_Sp, Pl_exp, Pl_year, ncol = 3)
```



В чем проблема?

Нет ли коллинеарности

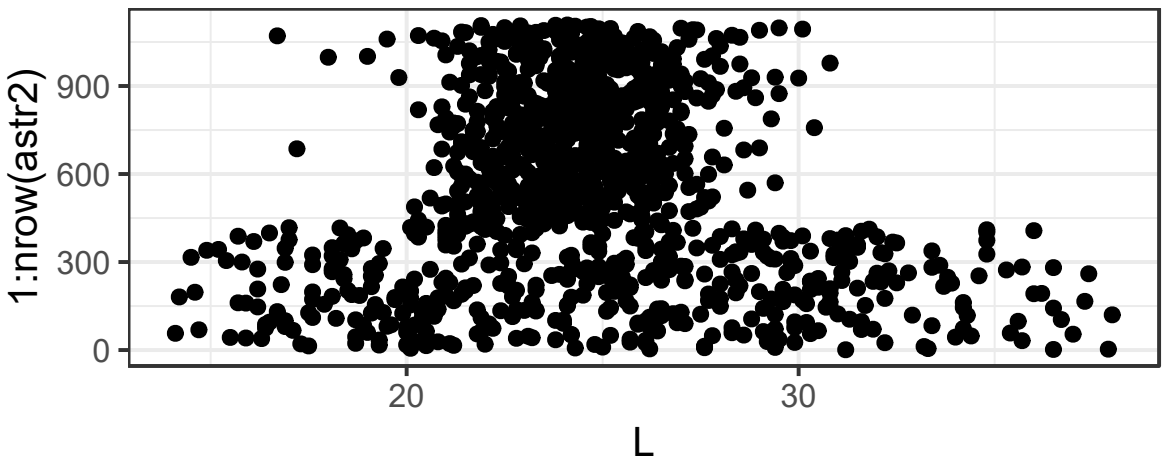
```
library(cowplot); library(ggplot2); theme_set(theme_bw())

Pl_Sp <- ggplot(astr2, aes(x = Sp, y = L)) + geom_boxplot()
Pl_exp <- ggplot(astr2, aes(x = Experiment, y = L)) + geom_boxplot()
Pl_year <- ggplot(astr2, aes(x = Year, y = L)) + geom_boxplot()
plot_grid(Pl_Sp, Pl_exp, Pl_year, ncol = 3)
```



В чем проблема? Размер распределен более-менее равномерно между градациями, коллинеарности нет. Однако в 2016 году была проведена более жесткая сортировка по размеру.

Есть ли выбросы?

```
ggplot(astr2, aes(y = 1:nrow(astr2))) + geom_point(aes(x = L) )
```



Выбросов нет, но проблема разного разброса в разные годы опять видна. В модель обязательно надо включить взаимодействие `Year:L`

Подбираем модель

Компоненты модели, котоую мы строим

Фиксированная часть модели

Предикторы:

- `Sp` - дискретный фактор (градации: `Tr`, `Ed`)
- `Year` - дискретный фактор (градации: `2015`, `2016`)
- `L` - непрерывный предиктор
- Взаимодействия первого порядка `Sp:Year`, `Sp:L`, `L:Year`
- Взаимодействия второго порядка `Sp:Year:L`

Случайная часть модели

- Эффект эксперимента
- Эффект контейнера в пределах эксперимента
- Остатки

Модель, которую мы строим

$Out_i \sim Binomial(n = 1, \pi_i)$

$E(y_i) = \pi_i$

$ln(\frac{\pi_i}{1 - \pi_i}) = \eta_i$ — функция связи логит, переводит вероятности в логиты.

$$\boldsymbol{\eta}_{ijk} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Sd} + \mathbf{Zb}$$

где i – наблюдение, jk – j-й контейнер в k-м эксперименте, k – эксперимент.

X - модельная матрица, описывающая фиксированные предикторы.

S - матрица, описывающая распределение испытаний по контейнерам.

Z - матрица, описывающая распределение испытаний по экспериментам.

$\mathbf{d}_{jk}$ - дисперсия, связанная с контейнерами в пределах эксперимента.

$\mathbf{b}_k$ - дисперсия, связанная с экспериментами.

Модель со случайным свободным членом (random intercept model)

```
library(lme4)
model1_ri <- glmer(Out ~ L*Sp*Year +
                     (1|Experiment/Box) , data = astr2,
                   family = binomial(link = "logit"))
```

```
# Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
# failed to converge with max|grad| = 0.00386958 (tol = 0.001, component 1)
```

Мы получили предупреждение (Warning), что модель не сошлась.

Если модель не сходится

Это обычная история для функции `glmer()`, в которой реализуются очень сложные числовые итерационные решения.

Также справку по этой проблеме можно посмотреть, набрав

```
?convergence
```

Если модель не сходится

Для исправления можно пойти по нескольким путями

1. В первую очередь надо проверить нет ли каких-то проблем с данными (ошибки в набивке)
2. Часто проблему можно решить за счет стандартизации непрерывных предикторов.
3. Можно настроить вычислительный алгоритм функции `glmer()`. Потребуется советов программистов и не всегда сработает (универсального решения нет). Про настройку алгоритма можно почитать на сайте https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html или на форуме https://stackoverflow.com/
4. Можно подобрать модель с помощью другого пакета (помимо `lme4`, GLMM реализованы еще в пакетах `glmmADMB`, `MASS`, `glmmML`), но там могут всплыть свои проблемы и появиться новые ограничения.
5. Можно упростить модель (не рассматривать, если возможно, взаимодействия предикторов, или убрать из модели взаимодействия высоких порядков).

Стандартизация непрерывного предиктора

```
astr2$L_scaled <- as.numeric(scale(astr2$L))

model1_ri <- glmer(Out ~ L_scaled*Sp*Year +
                     (1|Experiment/Box) , data = astr2,
                   family = binomial(link = "logit"))
```

Все посчиталось.

Модель со случайным свободным членом и случайным угловым коэффициентом (random intercept and random slope model)

Характер связи зависимой переменной с предикторами может меняться от контейнера к контейнеру и от эксперимента к эксперименту. Мы должны проверить не улучшит ли это модель.

```
model1_rsi_1<- glmer(Out ~ L_scaled * Sp * Year + (1 + Sp|Experiment/Box) ,
                data = astr2, family = binomial(link = "logit"))

# Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
# failed to converge with max|grad| = 0.00343151 (tol = 0.001, component 1)
```

```
model1_rsi_2 <- glmer(Out ~ L_scaled * Sp * Year + (1 + L_scaled |Experiment/
            data = astr2, family = binomial(link = "logit"))

# Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
# failed to converge with max|grad| = 0.0174684 (tol = 0.001, component 1)
```

Обе модели не сошлись.

Настраиваем алгоритм функции glmer()

В данном случае мы увеличили количество итераций.

```
model1_rsi_1<- glmer(Out ~ L_scaled * Sp * Year + (1 + Sp|Experiment/Box) ,
               data = astr2, family = binomial(link = "logit"),
               control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=2e

model1_rsi_2 <- glmer(Out ~ L_scaled * Sp * Year + (1 + L_scaled |Experiment/
            data = astr2, family = binomial(link = "logit"),
            control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=2e5))
```

Сработало. Но не факт, что сработает в других моделях.

Сравниваем три модели

Сравниваем три модели

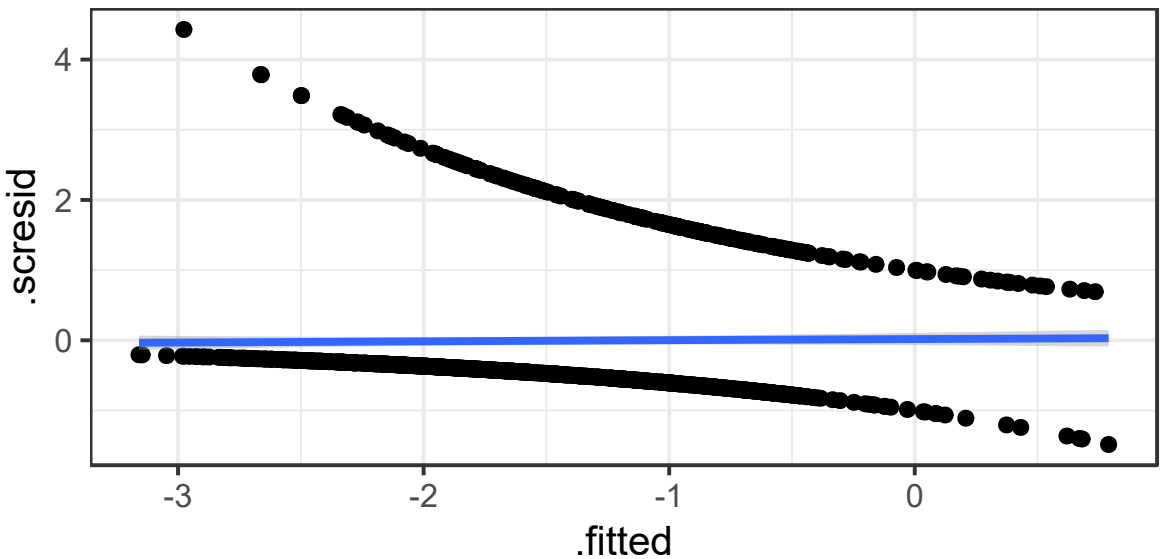
```
AIC(model1_ri, model1_rsi_1, model1_rsi_2)
```

```
#                df      AIC
# model1_ri      10 1081.682
# model1_rsi_1   14 1086.690
# model1_rsi_2   14 1087.618
```

Останавливаем выбор на модели со случайным отрезком `model1_ri`

Диагностика модели: линейность связи

```
library(ggplot2)
model1_diagn <- fortify(model1_ri)
ggplot(model1_diagn, aes(x = .fitted, y = .scresid)) + geom_point() + geom_sm
```



Нарушений линейности нет

Диагностика модели: избыточность дисперсии

```
library(sjstats)
overdisp(model1_ri)
```

```
#
# # Overdispersion test
#
#        dispersion ratio = 0.9656
#   Pearson's Chi-Squared = 1060.2624
#                 p-value = 0.7883
```

Избыточной дисперсии нет

Дорабатываем модель

Задание: Проведите упрощение модели в соответствии с протоколом backward selection

Задание: Проведите упрощение модели в соответствии с протоколом backward selection

```
drop1(model1_ri)

# Single term deletions
#
# Model:
# Out ~ L_scaled * Sp * Year + (1 | Experiment/Box)
#                 Df    AIC
# <none>              1081.7
# L_scaled:Sp:Year  1 1081.5
```

```
model2 <- update(model1_ri, .~.-L_scaled:Sp:Year)
```

Упроаем модель: шаг 2.

```
drop1(model2)
```

```
# Single term deletions
#
# Model:
# Out ~ L_scaled + Sp + Year + (1 | Experiment/Box) + L_scaled:Sp +
#     L_scaled:Year + Sp:Year
#               Df    AIC
# <none>           1081.5
# L_scaled:Sp    1 1079.5
# L_scaled:Year  1 1079.5
# Sp:Year        1 1079.5
```

```
model3 <- update(model2, . ~ . - L_scaled:Year)
```

Упроаем модель: шаг 3.

```
drop1(model3)
```

```
# Single term deletions
#
# Model:
# Out ~ L_scaled + Sp + Year + (1 | Experiment/Box) + L_scaled:Sp +
#     Sp:Year
#             Df    AIC
# <none>         1079.5
# L_scaled:Sp  1 1077.5
# Sp:Year      1 1077.5
```

```
model4 <- update(model3, . ~ . - L_scaled:Sp)
```

Упрощаем модель: шаг 4.

```
drop1(model4)
```

```
# Single term deletions
#
# Model:
# Out ~ L_scaled + Sp + Year + (1 | Experiment/Box) + Sp:Year
#          Df    AIC
# <none>      1077.5
# L_scaled  1 1107.5
# Sp:Year   1 1075.5
```

```
model5 <- update(model4, . ~ . - Sp:Year)
```

Упрощаем модель: шаг 5.

```
drop1(model5)
```

```
# Single term deletions
#
# Model:
# Out ~ L_scaled + Sp + Year + (1 | Experiment/Box)
#          Df    AIC
# <none>      1075.5
# L_scaled  1 1105.8
# Sp        1 1120.7
# Year      1 1074.1
```

```
model6 <- update(model5, . ~ . - Year)
```

Финальная модель

```
drop1(model6)
```

```
# Single term deletions
#
# Model:
# Out ~ L_scaled + Sp + (1 | Experiment/Box)
#          Df    AIC
# <none>      1074.1
# L_scaled  1 1104.4
# Sp        1 1119.5
```

Больше ничего удалить нельзя.

`model6` – финальная модель

Начальная и финальная модель

```
AIC(model1_ri, model6)
```

```
#           df      AIC
# model1_ri 10 1081.682
# model6     5 1074.137
```

Упрощение модели не привело к ее ухудшению. Мы удалили из фиксированной части модели избыточные предикторы и взаимодействия.

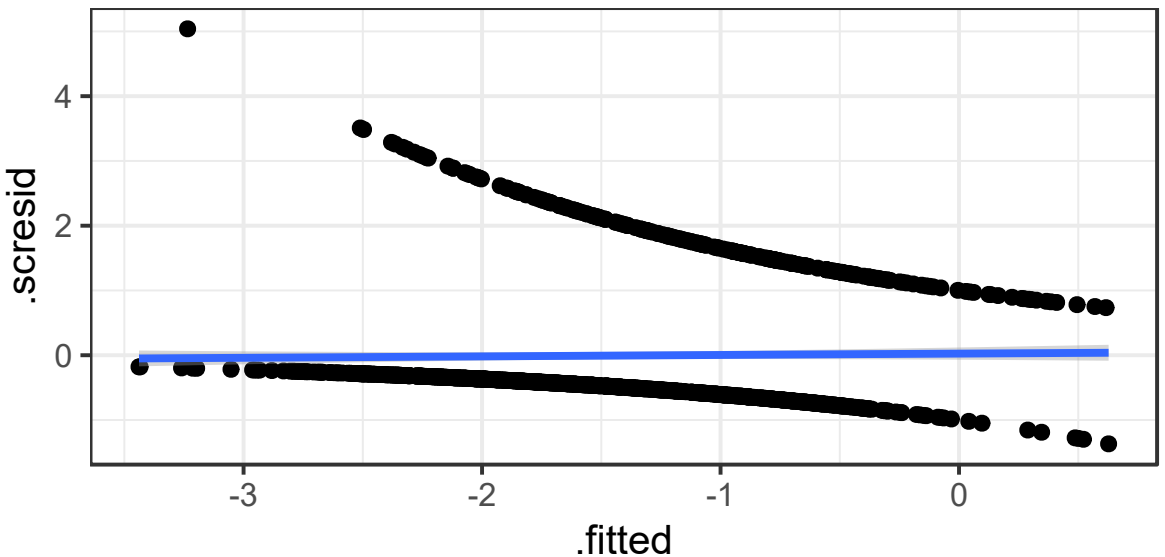
В фиксированной части модели осталось только два предиктора

```
summary(model6)$call
```

```
# glmer(formula = Out ~ L_scaled + Sp + (1 | Experiment/Box), data = astr2,
#     family = binomial(link = "logit"))
```

Диагностика финальной модели: линейность связи

```
model6_diagn <- fortify(model6)
ggplot(model6_diagn, aes(x = .fitted, y = .scresid)) + geom_point() + geom_sm
```



Нарушений линейности нет

Диагностика финальной модели: избыточность дисперсии

```
overdisp(model6)
```

```
#
# # Overdispersion test
#
#        dispersion ratio = 0.9596
#   Pearson's Chi-Squared = 1058.4863
#                 p-value = 0.8281
```

Избыточной дисперсии нет

Первые итоги подбора модели

В фиксированной части финальной модели осталось только два пердиктора (`L_scaled` и `Sp`), не взаимодействующих друг с другом. То есть выбор звездами того или иного вида жертв не зависит от их размера.

Важно: в финальную модель не вошел фактор `Year`. Это означает, что в разные годы характер связи переменной отклика с предикторами оставался одним и тем же.

Смотрим в summary()

Что вы можете сказать о случайных и фиксированных эффектах?

```
summary(model6)
```

```
# Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod
# ]
#  Family: binomial  ( logit )
# Formula: Out ~ L_scaled + Sp + (1 | Experiment/Box)
#    Data: astr2
#
#      AIC      BIC   logLik deviance df.resid
#   1074.1   1099.2   -532.1   1064.1     1103
#
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -1.3684 -0.5339 -0.3925 -0.2817  5.0397
#
# Random effects:
#  Groups         Name        Variance Std.Dev.
#  Box:Experiment (Intercept) 0.08148  0.2854
#  Experiment     (Intercept) 0.07328  0.2707
# Number of obs: 1108, groups:  Box:Experiment, 41; Experiment, 4
#
# Fixed effects:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept) -1.87617    0.18673 -10.047  < 2e-16 ***
# L_scaled    -0.44467    0.08155  -5.453 4.96e-08 ***
# SpTr         1.05869    0.15626   6.775 1.24e-11 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Correlation of Fixed Effects:
```

Какова роль фиксированных предикторов?

Мы получили результаты, говорящие о значимости связи отклика с фиксированными предикторами

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.87617    0.18673 -10.047  < 2e-16 ***
L_scaled    -0.44467    0.08155  -5.453 4.96e-08 ***
SpTr         1.05869    0.15626   6.775 1.24e-11 ***
```

Дайте трактовку полученным коэффициентам

Какова роль фиксированных предикторов?

Мы получили результаты, говорящие о значимости связи отклика с фиксированными предикторами

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.87617    0.18673 -10.047  < 2e-16 ***
L_scaled    -0.44467    0.08155  -5.453 4.96e-08 ***
SpTr         1.05869    0.15626   6.775 1.24e-11 ***
```

Дайте трактовку полученным коэффициентам

Отношения шансов быть съеденной изменяются в $e^{-0.44467}$ = 0.6 раз при изменении стандартизированной длины на единицу.

Отношения шансов быть съеденной у мидии из группы Tr (вид-вселенец) выше, чем у мидии из группы Ed (коренной вид), в $e^{1.05869}$ = 2.9 раза.

Проблема ICC для GLMM?

Внутриклассовая корреляция (intraclass correlation coefficient, ICC) очень удобный показатель для оценки роли случайных эффектов.

Чем ниже ICC, тем ниже роль группирующих факторов

ICC можно интерпретировать, как долю изменчивости объясненной наличием группирующего фактора.

В случае GLMM (при биномиальном или пуассоновском распределении отклика) величина дисперсии закономерно связана с матожиданием, следовательно она изменяется вместе с изменением предикторов, входящих в фиксированную часть модели. Это не позволяет вычислить напрямую, как мы это делали в случае LMM.

Приблизительные значения ICC

Для биномиальных GLMM приблизительную оценку ICC можно получить с помощью функции `icc()` из пакета `sjstats`

```
icc(model6)
```

```
#
# Intraclass Correlation Coefficient for Generalized linear mixed model
#
# Family : binomial (logit)
# Formula: Out ~ L_scaled + Sp + (1 | Experiment/Box)
#
#   ICC (Box:Experiment): 0.0237
#       ICC (Experiment): 0.0213
```

Все ICC очень небольшие, следовательно роль пространственных различий (разные контейнеры в пределах эксперимента) и временных различий (разные эксперименты) невелика.

Следовательно мы воспроизводили более или менее стандартные условия, которые не менялись при повторении экспериментов.

Визуализация модели

Два способа визуализаци

Описание роли случайных эффектов не является задачей исследования. Нас интересует характер связи отклика с фиксированными предикторами.

В фокусе исследования был вопрос о различиях в вероятности быть съеденной между двумя видами мидий, однако в финальной модели осталась и непрерывная ковариата.

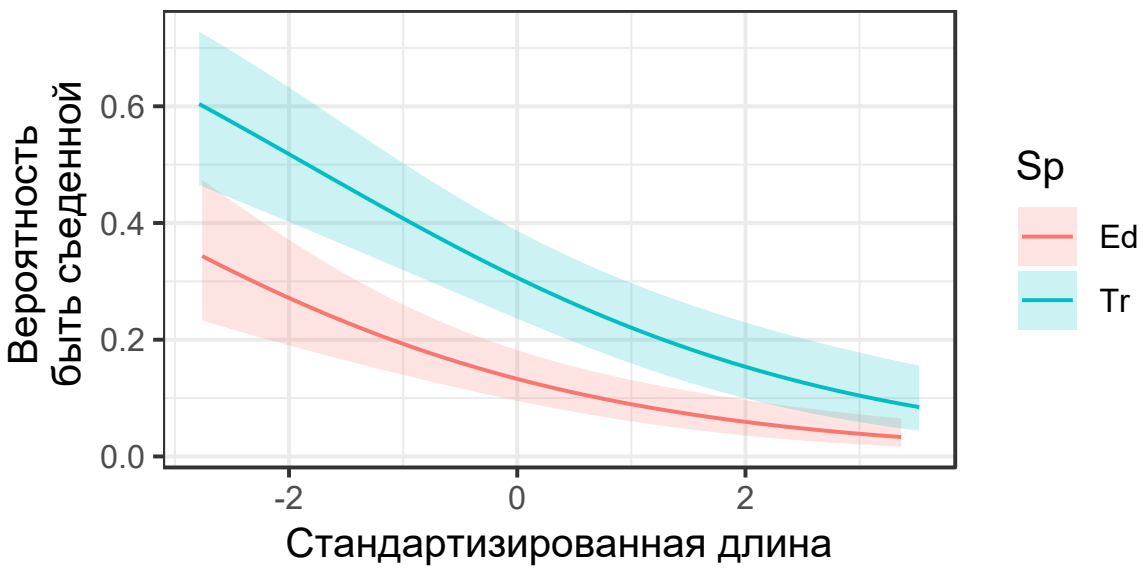
Визуализировать модель можно двумя способами: в виде логистических кривых и в виде столбчатой диаграммы.

Подготовка к визуализации в виде логистических кривых

```
logit_back <- function(x) exp(x)/(1 + exp(x)) # обратная логит-трансформация

library(dplyr)
new_data <- astr2 %>% group_by(Sp) %>%
  do(data.frame(L_scaled = seq(min(.$L_scaled),max(.$L_scaled),
                               length.out = 100)))

X <- model.matrix(~  L_scaled + Sp, data = new_data)
b <- fixef(model6)

new_data$fit_eta <- X %*% b
new_data$se_eta <- sqrt(diag(X %*% vcov(model6) %*% t(X)))

new_data$fit_pi <- logit_back(new_data$fit_eta)
new_data$lwr <- logit_back(new_data$fit_eta - 2 * new_data$se_eta)
new_data$upr <- logit_back(new_data$fit_eta + 2 * new_data$se_eta)
```

Логистические кривые

```
Pl_log <- ggplot(new_data, aes(x = L_scaled, y = fit_pi)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr, fill = Sp), alpha = 0.2) +
    geom_line(aes(color = Sp)) +
  labs(x = "Стандартизированная длина", y = "Вероятность \n быть съеденной" )
Pl_log
```



Согласно модели, чем больше длина мидии, тем меньше вероятность быть съеденной. Вероятность быть съеденной у мидий из группы Tr выше, чем у мидий из группы Ed.

Задание: Визуализируйте модель в виде столбчатой диаграммы

В фокусе исследования было сравнение вероятности быть съеденной у мидий двух групп: `Tr` vs `Ed`, а `L_scaled` - лишь ковариата, влияние которой нас может не интересовать.

Эффект влияния дискретного предиктора лучше отразить в виде столбчатой диаграммы, отражающей предсказание модели при среднем значении ковариаты.

Задание: Визуализируйте модель в виде столбчатой диаграммы

В фокусе исследования было сравнение вероятности быть съеденной у мидий двух групп: `Tr` vs `Ed`, а `L_scaled` - лишь ковариата, влияние которой нас может не интересовать.

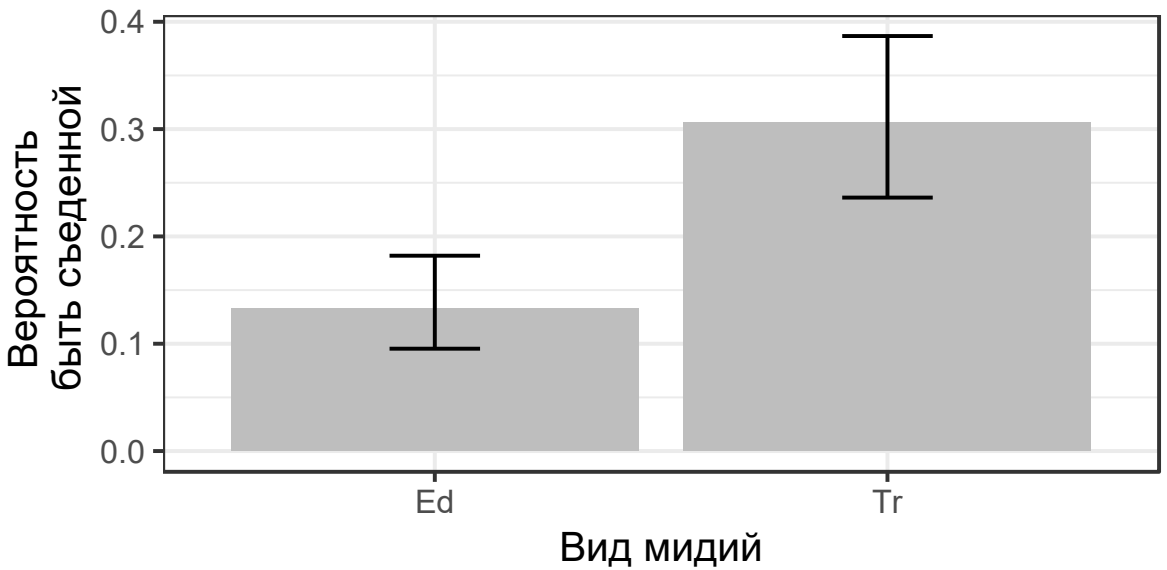
Эффект влияния дискретного предиктора лучше отразить в виде столбчатой диаграммы, отражающей предсказание модели при среднем значении ковариаты.

Ковариата стандартизована, ее среднее равно нулю.

```
new_data <- data.frame(Sp = c("Tr", "Ed"), L_scaled = 0)

X <- model.matrix(~  L_scaled + Sp, data = new_data)
b <- fixef(model6)
new_data$fit_eta <- X %*% b
new_data$se_eta <- sqrt(diag(X %*% vcov(model6) %*% t(X)))
new_data$fit_pi <- logit_back(new_data$fit_eta)
new_data$lwr <- logit_back(new_data$fit_eta - 2 * new_data$se_eta)
new_data$upr <- logit_back(new_data$fit_eta + 2 * new_data$se_eta)
```

Столбчатая диаграмма

```
ggplot(new_data, aes(x = Sp, y = fit_pi)) +
  geom_col(fill = "gray") +
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.2) +
  labs(x = "Вид мидий", y = "Вероятность \n быть съеденной" )
```



Дополнительные штрихи к модели

Проблема представления первичных данных при визуализации модели

На графике, визуализирующем модель, полезно приводить первичные данные. Если модель хороша, то первичные данные группируются вокруг соответствующей линии регрессии.

Однако переменная отклик в нашем случае принимает значение 1 или 0. Такие данные трудно визуализировать напрямую.

Можно отразить долю съеденных моллюсков каждого вида среди особей разных размерных классов. То есть отразить поведение некоторых усредненных данных. Это не первичные данные в прямом смысле.

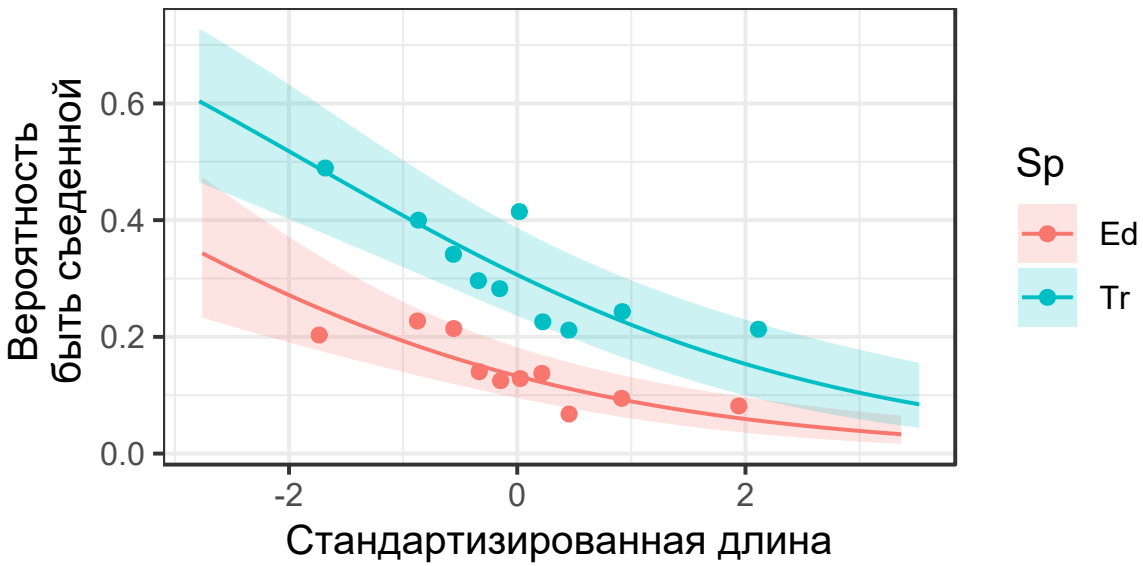
Разбиваем на размерные классы приблизительно рвного объема

```
astr2$Size_class <- ntile(astr2$L_scaled, 10)
table(astr2$Size_class, astr2$Sp)
```

```
#
#       Ed Tr
#    1  64 47
#    2  66 45
#    3  70 41
#    4  57 54
#    5  64 46
#    6  70 41
#    7  80 31
#    8  59 52
#    9  74 37
#    10 49 61
```

Средние показатели в каждом из размерных классов

```
Mean_Out <- astr2 %>% group_by(Size_class, Sp) %>%
  do(data.frame(Out = mean(.$Out), L_scaled = mean(.$L_scaled)))
Pl_log + geom_point(data = Mean_Out, aes(x = L_scaled, y = Out, color = Sp))
```



Точки отражают долю съеденных моллюсков в пределах каждого вида среди особей одинаковых размеров (без учета контейнера и эксперимента)

Еще один штрих: тестовая выборка

Гипотезу о том, что хищники с большей вероятностью атакуют мидий-вселенцев, мы подвергли суровому испытанию, проведя многочисленные дополнительные исследования.

Дополнительной проверкой будет работоспособность модели на данных, не включенных в анализ.

Тестовая выборка

В идеале тестовая выборка должна быть собрана специально по той же методике, что и выборка, лежащая в основе модели. Или тестовую выборку можно получить путем случайного разделения данных на обучающую и тестирующую части (кросс-валидация).

Если модель работает, то она должна давать предсказания близкие к наблюдаемым. Посмотрим, произойдет ли это если мы используем данные пилотного эксперимента, которые мы не использовали при подборе модели.

Важно Это учебный пример. В реальных исследованиях лучше не применять данные, которые уже были использованы для формулировок каких-то гипотез, в повторном анализе.

```
astr_test <- read.csv('data/aster_mussel.csv', header = TRUE)
```

Переподберем модель для нестандартизированной ковариаты

В тестовой выборке мидии имеют другие размеры, нежели в наборе данных, на которых была построена модель. Стандартизованная ковариата не годится.

```
model6_unscaled <- glmer(Out ~ L + Sp +
                    (1|Experiment/Box) , data = astr2,
                    family = binomial(link = "logit"))
```

Ничего не изменилось по сути

```
summary(model6)
```

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.87617    0.18673 -10.047  < 2e-16 ***
L_scaled    -0.44467    0.08155  -5.453 4.96e-08 ***
SpTr         1.05869    0.15626   6.775 1.24e-11 ***
```

```
summary(model6_unscaled)
```

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.01862    0.53776   1.894   0.0582 .
L           -0.11740    0.02153  -5.453 4.96e-08 ***
SpTr         1.05871    0.15626   6.775 1.24e-11 ***
```

Предсказания для новых данных

Задание: Сделайте предсказания для новых данных

Предсказания для новых данных

Задание: Сделайте предсказания для новых данных

```
X <- model.matrix(~ L + Sp, data = astr_test)
b <- fixef(model6_unscaled)
astr_test$predicted_pi <- logit_back(X %*% b)
```

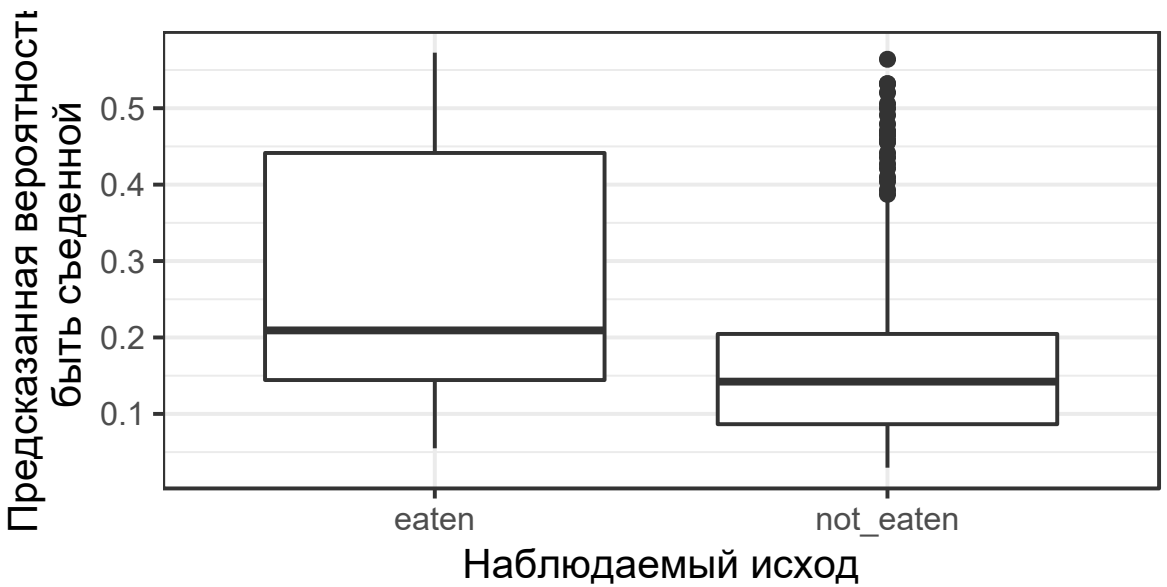
Визуализация связи наблюдений и предсказаний.

Задание: Предложите способ визуализировать соотношение предсказанных и наблюдаемых значений.

Визуализация связи наблюдений и предсказаний.

Задание: Предложите способ визуализировать соотношение предсказанных и наблюдаемых значений.

```
ggplot(astr_test, aes(x = Outcome, y = predicted_pi)) +
  geom_boxplot() + labs(x = "Наблюдаемый исход",
                        y = "Предсказанная вероятность \n быть съеденной")
```



Для реально съеденных мидий предсказанная вероятность быть съеденной, в среднем, выше, чем для особей, на которых звезды не напали.

Summary: Что мы знаем

Общие черты GLM и GLMM для бинарных откликов

В целом идеи лежащие в основе анализа аналогичны идеям GLM для бинарных откликов:

- В основе анализа стоит подбор параметров логистической регрессионной модели.
- Параметры логистической регрессии подбираются методом максимального правдоподобия.
- Угловые коэффициенты логистической регрессии позволяют сказать во сколько раз изменяется соотношение шансов для события при увеличении предиктора на единицу (или при переходе от базового уровня фактора к данному уровню).
- Для визуализации результатов лучше проводить обратное логит-преобразование и отражать зависимую переменную в терминах вероятностей.

Особенности GLMM для бинарных откликов

- Роль случайных факторов в GLMM в основе такая же, как и в случае с LMM. Группирующие факторы могут определять дисперсию свободного члена модели или дисперсию углового коэффициента и свободного члена модели.
- GLMM требует очень больших вычислительных ресурсов и далеко не всегда параметры модели легко вычисляются, часто модели не сходятся.
- Для лучшего схождения моделей их не надо чрезмерно усложнять.
- Часто помогает стандартизация непрерывных предикторов.
- Внутриклассовые корреляции могут быть вычислены лишь приблизительно, но их рассмотрение дает важную информацию. Если ICC равен нулю, то лучше отказаться от рассмотрения случайных факторов и применить обычную GLM.

Что почитать

- Zuur, A.F. et al. 2009. Mixed effects models and extensions in ecology with R. - Statistics for biology and health. Springer, New York, NY.