

Обобщенные линейные модели с бинарным откликом

Линейные модели...

Вадим Хайтов, Марина Варфоломеева

Кафедра Зоологии беспозвоночных, Биологический факультет, СПбГУ



Бинарные переменные вокруг нас



Числа и События

До сих пор в качестве переменной отклика мы рассматривали числовые данные.

- ▶ Содержание сухого вещества в икре
- ▶ Вес младенцев
- ▶ Объем легких
- ▶ Число посещений цветков опылителями.

А как быть, если мы хотим проанализировать связь появления того или иного **события** (произошло или не произошло) с некоторыми предикторами?

События и предикторы

В зависимости от предикторов события могут происходить чаще или реже – логика, совпадающая с логикой связи количественной переменной отклика с набором предикторов.

Например, по мере роста температуры воздуха летом чаще будут встречаться люди в шортах: событие “встретился человек в шортах” положительно связано с температурой воздуха.

Событие “покупка автомобиля” явно связана с предиктором “количество денег на счете”, однако эта связь может быть совсем непростой.

Бинарные данные вокруг нас

Бинарные данные – очень распространенный тип зависимых переменных

- ▶ Кто-то в результате лечебных процедур выжил или умер
- ▶ Обследованное животное заражено паразитами или здорово
- ▶ Футбольная команда выиграла или проиграла
- ▶ Блюдо вкусное или невкусное

Все эти события могут быть связаны с самыми разными предикторами и эту связь можно описать с помощью регрессионных моделей.

Обобщенные линейные модели позволяют моделировать в том числе и бинарные данные.

Пример – морские звезды и мидии

Различают ли морские звезды два вида мидий?

Атлантические мидии (*Mytilus edulis*) коренной для Белого моря вид, но недавно туда вселились мидии другого вида – тихоокеанские мидии (*M. trossulus*).



Вселенец имеет меньшую промысловую значимость и потенциально может влиять на структуру экосистемы. Важно понять, что регулирует их численность. Наиболее значимый фактор – это морские звезды, питающиеся мидиями.

- ▶ Различают ли морские звезды два вида мидий?
- ▶ Различают ли хищники мидий разных размеров?

Данные: Khalitov et al, 2018

Тонкости дизайна эксперимента

Морских звезд вместе с мидиями двух видов сажали в контейнеры. Через четыре дня совместного существования с хищником регистрировали состояние мидий.



Зависимая переменная:

- Outcome – состояние мидий (“eaten” – съедена, “not_eaten” – живая)

Предикторы в фокусе исследования:

- Sp – вид мидий (“Ed” – коренной вид, “Tr” – вселенец),
- L – размер мидий (мм).

Чего не хватает?

Как быть с контейнерами?

В этом эксперименте, помимо интересующих нас дискретного фактора S_p (вид мидии) и непрерывного предиктора L (размер), есть еще один фактор V_{ox} .

Этот фактор нас не интересует, но его нельзя не учитывать.

Мы должны включить в модель переменную V_{ox} в качестве дискретного фактора с 4 уровнями.

В лекциях, посвященных **смешанным линейным моделям**, мы научим вас, как включать в модель подобные факторы более правильным способом.

Читаем данные

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

```
#   Box    L Sp Outcome
# 1    1 33.1 Tr not_eaten
# 2    1 24.8 Ed not_eaten
# 3    1 33.0 Ed not_eaten
# 4    1 18.8 Ed not_eaten
# 5    1 34.0 Ed not_eaten
# 6    1 22.6 Ed not_eaten
```

Номер экспериментального контейнера закодирован числами, поэтому превращаем его в фактор.

```
astr$Box <- factor(astr$Box)
```

Знакомимся с данными

Нет ли пропущенных значений?

```
colSums(is.na(astr))
```

```
#      Box      L      Sp Outcome  
#       0       0       0         0
```

Каковы объемы выборок?

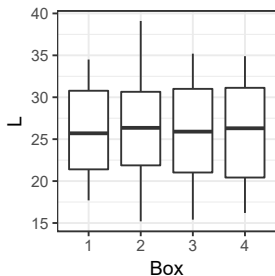
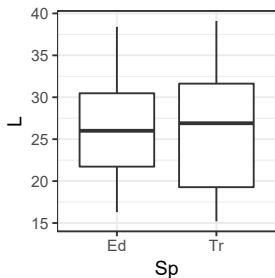
```
table(astr$Box)
```

```
#  
#  1  2  3  4  
# 66 68 74 78
```

Нет ли коллинеарности

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

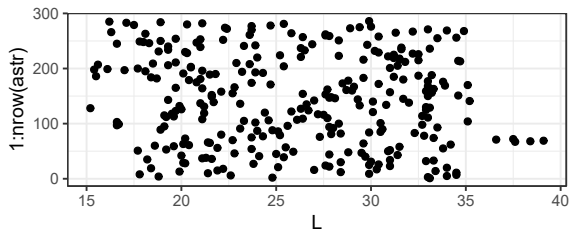
```
Pl_Sp <- ggplot(astr, aes(x = Sp, y = L)) + geom_boxplot()  
Pl_Box <- ggplot(astr, aes(x = Box, y = L)) + geom_boxplot()  
plot_grid(Pl_Sp, Pl_Box, ncol = 2)
```



Размер распределен
более-менее равномерно.
Коллинеарности нет.

Есть ли выбросы?

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



Выбросов нет.

Кодирование бинарной переменной

До сих пор зависимая переменная была числом,
а в данном случае Outcome – это текстовая переменная.

Бинарную переменную надо перекодировать в виде нулей и единиц:

- ▶ 1 – мидию съели,
- ▶ 0 – мидию не съели.

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

Простой линейной регрессией не обойтись

Что мы хотим построить?

Наша задача – построить модель, описывающую связь между переменной-откликом (съедена мидия или нет) и тремя предикторами: Sp , L и Box

Если бы мы строили GLM с нормальным распределением отклика, то она имела бы следующий вид:

$$Out_i \sim N(\mu_i, \sigma)$$

$$E(Out_i) = \mu_i$$

$\mu_i = \eta_i$ – функция связи “идентичность”

$$\eta_i = b_0 + b_1 Sp_{Tri} + b_2 L_i + b_3 Box_{2i} + b_3 Box_{3i} + Interactions,$$

где *Interactions* — это все взаимодействия.

Лобовая атака?

Строим модель по накатанной дороге.

Мы не знаем, взаимодействуют ли дискретные факторы Box, Sp и непрерывный предиктор L.

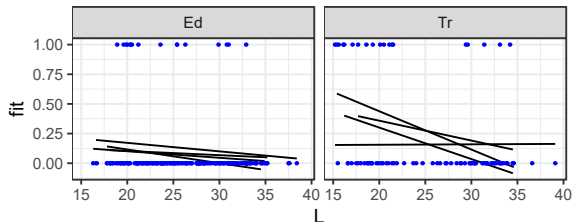
В полной модели мы должны учесть как влияние самих предикторов, так и влияние их взаимодействия.

```
mod_norm <- glm(Out ~ Sp * L * Box, data = astr)
```

Все посчиталось...

Посмотрим что получилось

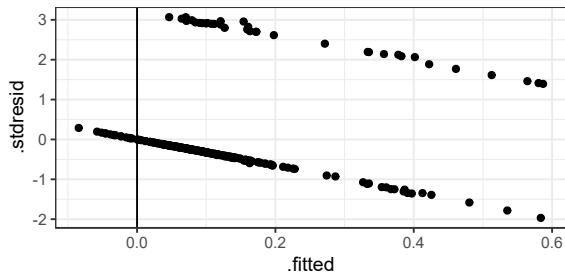
```
library(dplyr)
new_data <- astr %>% group_by(Sp, Box)%>%
  do(data.frame(L = seq(min(.$L), max(.$L), length.out = 100)))
new_data$fit <- predict(mod_norm, newdata = new_data) # Предсказанные значения
ggplot(new_data, aes(x = L, y = fit)) +
  geom_line(aes(group = Box)) + facet_wrap(~ Sp, ncol = 2) +
  geom_point(data = astr, aes(x = L, y = Out), size = 0.5, color = 'blue')
```



Во-первых, непонятно, что за величина отложена по оси OY.

Диагностика модели

```
mod_norm_diag <- fortify(mod_norm)
ggplot(mod_norm_diag, aes(x = .fitted, y = .stdresid)) +
  geom_point() + geom_vline(xintercept = 0)
```



Во-вторых, модель
предсказывает
отрицательные значения.
Простая линейная модель
категорически не годится!

Логистическая кривая



Бинарные данные можно представлять и иначе

Бинарные данные очень неудобны для работы. Вместо того, чтобы моделировать наличие нулей и единиц, мы будем моделировать вероятности получения единиц.

Появляется новое обозначение:

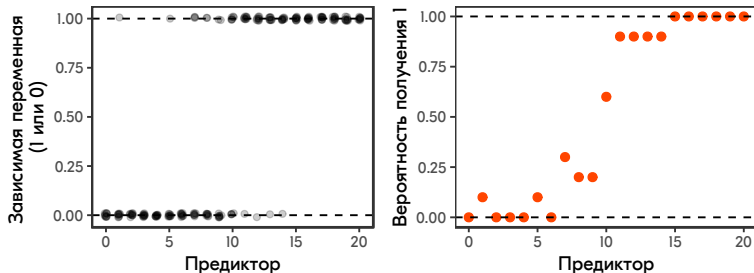
- ▶ π_i – вероятность события $y_i = 1$ при данных условиях,
- ▶ $1 - \pi_i$ – вероятность альтернативного события $y_i = 0$.

π_i – непрерывная величина, варьирующая от 0 до 1.

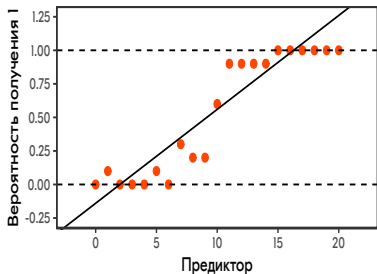
Симулированный пример: От дискретных значений к оценкам вероятностей

От 1 и 0 (слева) можно перейти к π_i – оценкам вероятности положительных исходов (справа).

Мы можем проиллюстрировать этот переход, изобразив доли в общем количестве исходов **при данном значении предиктора** $p_{y=1|x_i}$ (красные точки). И π_i и $p_{y=1|x_i}$ варьируют от 0 до 1.



Симулированный пример: Можно ли подобрать простую линейную регрессию?

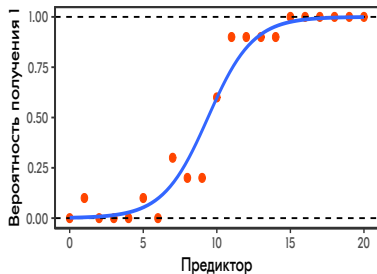


Связь зависимой переменной с предиктором можно было бы описать прямой:

$$\pi_i = \beta_0 + \beta_1 x_i$$

Но! Вероятность события, может принимать значения только от 0 до 1. А прямая линия ничем не ограничена и может выходить за пределы интервала $[0, 1]$.

Симулированный пример: Логистическая кривая

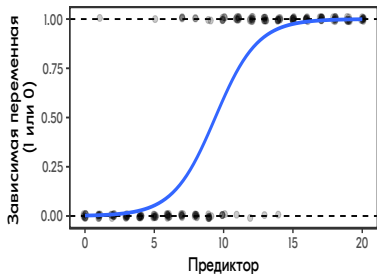


Связь вероятности положительного исхода и значений предиктора можно описать логистической кривой:

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

Логистическая кривая удобна для описания вероятностей, т.к. ее значения лежат в пределах от 0 до 1.

Симулированный пример: Логистическая кривая



В реальной жизни нам не потребуется даже рассчитывать доли положительных исходов от общего количества.

Благодаря GLM мы сможем оценить вероятности непосредственно по исходным данным.

Шансы и логиты

Шансы – еще один способ выразить бинарную переменную отклика

В обыденной речи мы часто используем фразы подобные такой:

“Шансы на победу 1 к 3”: в одном случае выигрыш в трех проигрыш

Шансы – это тоже оценка вероятности события. Шансы показывают сколько в данной системе положительных исходов и сколько отрицательных.

Отношение шансов

Шансы (odds) часто представляют в виде отношения шансов (odds ratio): $odds = \frac{n_+}{n_-}$

Если отношение шансов > 1 , то вероятность наступления события выше, чем вероятность того, что оно не произойдет. Если отношение шансов < 1 , то наоборот.

Если можно оценить вероятность положительного события, то отношение шансов выглядит так : $odds = \frac{\pi}{1-\pi}$

Отношение шансов варьирует от 0 до $+\infty$.

Отношение шансов

Если отношение шансов = 1, то вероятность того, что событие произойдет равно вероятности того, что событие не произойдет.

Асимметрия: отношение шансов от 1 до $+\infty$ говорит о том, что вероятность того, что событие произойдет, выше, чем вероятность того, что оно не произойдет, но если наоборот, то отношение шансов “зажато” между 0 и 1.

Логиты

Отношение шансов можно преобразовать в *Логиты* (logit):

$$\ln(odds) = \ln\left(\frac{\pi}{1 - \pi}\right)$$

Значения логитов – это трансформированные оценки вероятности события.

Логиты варьируют от $-\infty$ до $+\infty$.

Логиты симметричны относительно 0, т.е. $\ln(1)$.

Для построения моделей в качестве зависимой переменной удобнее брать логиты.

Немного алгебры: Логиты в качестве зависимой переменной

Докажем, что логит преобразование линеаризует логистическую кривую

Когда предиктор один, логистическая модель принимает такую форму:

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Обозначим для краткости $\beta_0 + \beta_1 x \equiv z$

Давайте докажем, что логит преобразование $\text{logit}(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right)$ сделает логистическую функцию линейной, т.е. что

$$\ln\left(\frac{\pi}{1 - \pi}\right) = z$$

Подставим выражение для π в формулу логита

$$\ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{\frac{e^z}{1+e^z}}{1-\frac{e^z}{1+e^z}}\right)$$

Логарифм отношения равен разности логарифмов, тогда:

$$\ln\left(\frac{e^z}{1+e^z}\right) - \ln\left(1 - \frac{e^z}{1+e^z}\right)$$

Вторую дробь можно упростить:

Подставим выражение для π в формулу логита

$$\ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{\frac{e^z}{1+e^z}}{1-\frac{e^z}{1+e^z}}\right)$$

Логарифм отношения равен разности логарифмов, тогда:

$$\ln\left(\frac{e^z}{1+e^z}\right) - \ln\left(1 - \frac{e^z}{1+e^z}\right)$$

Вторую дробь можно упростить:

$$\ln\left(\frac{e^z}{1+e^z}\right) - \ln\left(\frac{1+e^z-e^z}{1+e^z}\right) = \ln\left(\frac{e^z}{1+e^z}\right) - \ln\left(\frac{1}{1+e^z}\right)$$

Продолжаем преобразования

$$\ln\left(\frac{e^z}{1+e^z}\right) - \ln\left(\frac{1}{1+e^z}\right) = \ln(e^z) - \ln(1+e^z) - (\ln(1) - \ln(1+e^z))$$

Продолжаем преобразования

$$\ln\left(\frac{e^z}{1+e^z}\right) - \ln\left(\frac{1}{1+e^z}\right) = \ln(e^z) - \ln(1+e^z) - (\ln(1) - \ln(1+e^z))$$

$$\ln(e^z) - \ln(1+e^z) - 0 + \ln(1+e^z) = \ln(e^z) = z$$

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

Т.е. после логит-преобразования логистическая кривая становится прямой.

Связывающая функция (link function)

Мы уже знаем: Для линеаризации связи между предикторами и зависимой переменной применяется связывающая функция.

Функция логит-преобразования $g(E(y)) = \ln\left(\frac{\pi}{1-\pi}\right)$ это одна из возможных связывающих функций, применяемых для анализа бинарных переменных отклика.

Другие связывающие функции: probit, cloglog.

Логика математических преобразований

1. От дискретной оценки событий (1 или 0) переходим к оценке вероятностей.
2. Связь вероятностей с предиктором описывается логистической кривой.
3. Если при помощи функции связи перейти от вероятностей к логитам, то связь с предиктором будет описываться прямой линией.
4. Параметры линейной модели для такой прямой можно оценить при помощи линейной модели.

Теперь мы готовы сформулировать модель в математическом виде.

GLM с биномиальным распределением отклика

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

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

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

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

Чтобы перейти обратно от логитов к вероятностям, применяется логистическое преобразование (это функция, обратная функции связи):

$$\pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

Вернемся к морским звездам и мидиям

GLM с биномиальным распределением отклика

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

$$E(Out_i) = \pi_i$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

Полная модель в изучаемой системе включает много членов:

- ▶ Главные предикторы: Sp, L, Box
- ▶ Взаимодействия первого порядка: $Sp:L, Sp:Box, L:Box$
- ▶ Взаимодействия второго порядка: $Sp:L:Box$

```
mod <- glm(Out ~ Sp*L*Box, family = binomial(link = 'logit'), data = astr)
```

Анализ девиансы для полной модели

```
library(car)  
Anova(mod)
```

```
# Analysis of Deviance Table (Type II tests)  
#  
# Response: Out  
#  
#      LR Chisq Df Pr(>Chisq)  
# Sp      7.28  1  0.00696 **  
# L     12.35  1  0.00044 ***  
# Box      1.06  3  0.78597  
# Sp:L      0.05  1  0.82686  
# Sp:Box     2.57  3  0.46335  
# L:Box      2.06  3  0.56055  
# Sp:L:Box   4.86  3  0.18203  
# ---  
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Эту модель можно упростить!

Упрощение модели: Шаг 1

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

```
# Single term deletions
```

```
#  
# Model:  
# Out ~ Sp * L * Box  
#
```

	Df	Deviance	AIC	LRT	Pr(>Chi)
# <none>		180	212		
# Sp:L:Box	3	185	211	4.86	0.18

```
mod2 <- update(mod, . ~ . - Sp:L:Box)
```

Упрощение модели: Шаг 2

```
drop1(mod2, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# Out ~ Sp + L + Box + Sp:L + Sp:Box + L:Box
#           Df Deviance AIC    LRT Pr(>Chi)
# <none>                185 211
# Sp:L      1          185 209 0.048    0.83
# Sp:Box    3          187 207 2.567    0.46
# L:Box     3          187 207 2.058    0.56

mod3 <- update(mod2, . ~ . - Sp:L)
```

Упрощение модели: Шаг 3

```
drop1(mod3, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# Out ~ Sp + L + Box + Sp:Box + L:Box
#           Df Deviance AIC   LRT Pr(>Chi)
# <none>                185 209
# Sp:Box   3           188 206 2.65    0.45
# L:Box    3           187 205 2.27    0.52
```

```
mod4 <- update(mod3, . ~ . - L:Box)
```

Упрощение модели: Шаг 4

```
drop1(mod4, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# Out ~ Sp + L + Box + Sp:Box
#      Df Deviance AIC    LRT Pr(>Chi)
# <none>      187 205
# L      1      200 216 12.35  0.00044 ***
# Sp:Box 3      190 202  3.01  0.39031
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mod5 <- update(mod4, . ~ . - Sp:Box)
```

Упрощение модели: Шаг 5

```
drop1(mod5, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# Out ~ Sp + L + Box
#           Df Deviance AIC    LRT Pr(>Chi)
# <none>          190 202
# Sp           1      198 208   7.48  0.00625 **
# L            1      202 212  11.81  0.00059 ***
# Box          3      191 197   1.12  0.77165
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mod6 <- update(mod5, . ~ . - Box)
```

Упрощение модели: Шаг 6

```
drop1(mod6, test = 'Chi')
```

```
# Single term deletions
#
# Model:
# Out ~ Sp + L
#           Df Deviance AIC    LRT Pr(>Chi)
# <none>           191 197
# Sp         1       199 203   7.89   0.00496 **
# L          1       203 207  11.75   0.00061 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Больше никаких предикторов исключать нельзя: mod6 – финальная модель.

АІС для финальной модели

```
AIC(mod, mod2, mod3, mod4, mod5, mod6)
```

```
#      df AIC  
# mod  16 212  
# mod2 13 211  
# mod3 12 209  
# mod4  9 205  
# mod5  6 202  
# mod6  3 197
```

Информационный критерий Акайке показывает, что по мере удаления предикторов модель становится лучше.

Финальная модель (mod6) лучше, чем полная модель, с которой мы начинали.

Смысл коэффициентов в моделях с бинарной переменной отклика

Что за модель мы построили?

```
summary(mod6)
```

```
#
# Call:
# glm(formula = Out ~ Sp + L, family = binomial(link = "logit"),
#      data = astr)
#
# Deviance Residuals:
#      Min       1Q   Median       3Q      Max
# -1.058  -0.510  -0.410  -0.289   2.593
#
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)    0.399      0.875    0.46  0.6483
# SpTr           1.070      0.379    2.82  0.0047 **
# L             -0.113      0.035   -3.24  0.0012 **
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
#      Null deviance: 212.58  on 285  degrees of freedom
# Residual deviance: 191.24  on 283  degrees of freedom
# AIC: 197.2
#
# Number of Fisher Scoring iterations: 5
```

$$Out_i \sim \text{Binomial}(n = 1, \pi)$$

$$E(Out_i) = \pi_i$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

$$\eta_i = 0.399 + 1.07Sp_{Tr\,i} - 0.113L_i$$

Что означают коэффициенты модели?

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

$$E(Out_i) = \pi_i$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

$$\eta_i = 0.399 + 1.07Sp_{Tr_i} - 0.113L_i$$

- ▶ b_0 – интерсепт, логарифм отношения шансов для базового уровня дискретного фактора.
- ▶ b_1 – **на сколько единиц изменяется логарифм отношения шансов (logit)** для данного уровня (Tr) дискретного фактора Sp по сравнению с базовым уровнем (Ed).
- ▶ b_2 – **на сколько единиц изменяется логарифм отношения шансов (logit)**, если значение предиктора (L) изменяется на единицу.

Немного алгебры для понимания сути коэффициентов

Предположим, что у нас в модели есть только один непрерывный предиктор x .

Посмотрим, как изменятся предсказанные моделью значения, если значение непрерывного предиктора изменится на 1.

Мы знаем, что в терминах логитов модель выглядит вот так:

$$\eta = \ln\left(\frac{\pi}{1-\pi}\right) = \ln(odds)$$

Тогда разница между значениями η для $x + 1$ и x – это логарифм соотношения шансов при этих значениях предиктора:

$$\eta_{x+1} - \eta_x = \ln(odds_{x+1}) - \ln(odds_x) = \ln\left(\frac{odds_{x+1}}{odds_x}\right)$$

Продолжим преобразования

$$\ln\left(\frac{odds_{x+1}}{odds_x}\right) = b_0 + b_1(x + 1) - b_0 - b_1x = b_1$$

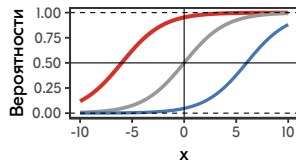
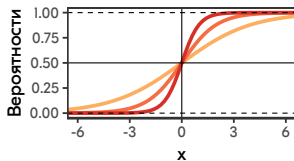
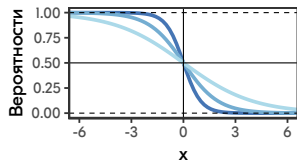
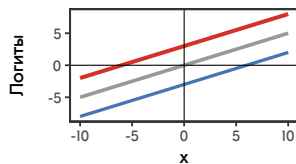
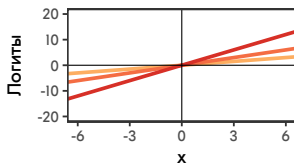
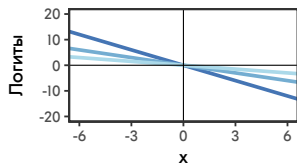
$$\ln\left(\frac{odds_{x+1}}{odds_x}\right) = b_1$$

$$\frac{odds_{x+1}}{odds_x} = e^{b_1}$$

Полученная величина e^{b_1} показывает, **во сколько раз изменится отношение шансов** при увеличении предиктора на единицу.

Для дискретных факторов e^{b_1} покажет, во сколько раз различается отношение шансов для данного уровня по сравнению с базовым.

Геометрическая интерпретация коэффициентов



Отрицательный
угловой
коэффициент

- $b_1 = -0.5$
- $b_1 = -1$
- $b_1 = -2$

Положительный
угловой
коэффициент

- $b_1 = 0.5$
- $b_1 = 1$
- $b_1 = 2$

Интерсепт

- $b_0 = 3$
- $b_0 = 0$
- $b_0 = -3$

Смысл интерсепта b_0

Величина e^{b_0} показывает отношение шансов для события, когда все предикторы равны нулю.

Когда предикторы физически не могут принимать нулевые значения, у этой величины нет смысла.

Но если произведена стандартизация предикторов, то смысл появится. У стандартизованных величин среднее значение равно нулю. Поэтому e^{b_0} покажет соотношение шансов для события при средних значениях предикторов.

Трактуем коэффициенты модели

$$\eta_i = 0.399 + 1.07I_{Tr,i} - 0.113L_i$$

- ▶ При увеличении длины тела мидии на 1 мм отношения шансов быть съеденной увеличатся в $e^{-0.113} = 0.893$ раза. То есть мидия, имеющая больший размер, имеет меньше шансов быть съеденной (больше шансов на выживание)
- ▶ Отношение шансов быть съеденной у мидии, относящиеся к группе Tr дискретного фактора Sp, в $e^{1.07} = 2.915$ раза выше, чем у мидии относящейся к базовому уровню (Ed). То есть вероятность выжить у мидии из группы Tr меньше, чем у мидии из группы Ed.

Диагностика модели с бинарным откликом

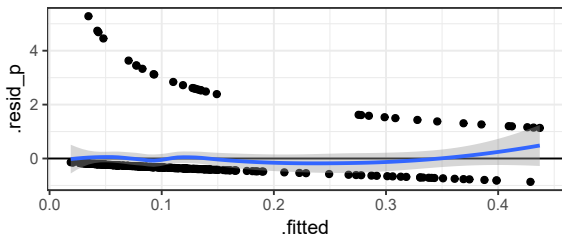
Условия применимости GLM с бинарной переменной-откликом

- ▶ Случайность и независимость наблюдений.
- ▶ Линейность связи переменной отклика с предиктором (с учетом связывающей функции).
- ▶ Отсутствие сверхдисперсии (форма связи среднего с дисперсией должна быть как у величины с биномиальным распределением).
- ▶ Отсутствие коллинеарности предикторов.

Линейность связи

Мы должны выяснить, нет ли криволинейного паттерна в остатках. Самый простой способ – это построить график остатков от предсказанных значений и наложить на него сглаживающую функцию, подобранную методом loess.

```
mod6_diag <- data.frame(.fitted = predict(mod6, type = 'response'),  
                        .resid_p = resid(mod6, type = 'pearson'))  
ggplot(mod6_diag, aes(y = .resid_p, x = .fitted)) + geom_point() +  
  geom_hline(yintercept = 0) + geom_smooth(method = 'loess')
```



Явного криволинейного паттерна нет.

Проверка на сверхдисперсию

Важное свойство биномиального распределения – это зависимость между матожиданием и дисперсией.

Мат.ожидание – $E(y_i) = \pi_i$

Дисперсия – $var(y_i) = \pi_i(1 - \pi_i)$

То есть в распределении остатков не должно наблюдаться сверхдисперсии (overdispersion).

Еще раз смотрим на результаты

```
summary(mod6)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.399	0.875	0.46	0.6483
SpTr	1.070	0.379	2.82	0.0047 **
L	-0.113	0.035	-3.24	0.0012 **

--

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 212.58 on 285 degrees of freedom
Residual deviance: 191.24 on 283 degrees of freedom
AIC: 197.2

Number of Fisher Scoring iterations: 5

Важная строчка

(Dispersion parameter for binomial family taken to be 1)

Проверка на сверхдисперсию

```
# Функция для проверки наличия сверхдисперсии в модели (автор Ben Bolker)
# http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html
overdisp_fun <- function(model) {
  rdf <- df.residual(model) # Число степеней свободы N - p
  rp <- residuals(model,type='pearson') # Пирсоновские остатки
  Pearson.chisq <- sum(rp^2) # Сумма квадратов остатков, подчиняется Хи-квадрат распределен
  pratio <- Pearson.chisq/rdf # Отношение суммы квадратов остатков к числу степеней свободы
  pval <- pchisq(Pearson.chisq, df=rdf, lower.tail=FALSE) # Уровень значимости
  c(chisq=Pearson.chisq, ratio=pratio, rdf=rdf, p=pval) # Вывод результатов
}
```

```
overdisp_fun(mod6)
```

```
# chisq ratio rdf p
# 294.25 1.04 283.00 0.31
```

Избыточной дисперсии
не выявлено.

Ben Bolker's glmmFAQ
<http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html>

Визуализация модели

Данные для предсказаний

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

Давайте получим предсказания при помощи операций с матрицами, чтобы своими глазами увидеть работу функции связи.

Предсказания модели при помощи операций с матрицами

```
# Модельная матрица и коэффициенты
X <- model.matrix(~ Sp + L, data = new_data)
b <- coef(mod6)
# Предсказанные значения и стандартные ошибки...
# ...в масштабе функции связи (логит)

new_data$fit_eta <- X %*% b
new_data$se_eta <- sqrt(diag(X %*% vcov(mod6) %*% t(X)))

# ...в масштабе отклика (применяем функцию, обратную функции связи)

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

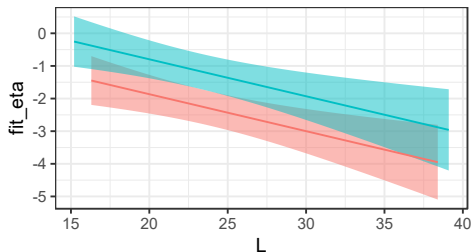
new_data$fit_pi <- logit_back(new_data$fit_eta)
new_data$lwr_pi <- logit_back(new_data$fit_eta - 2 * new_data$se_eta)
new_data$upr_pi <- logit_back(new_data$fit_eta + 2 * new_data$se_eta)

head(new_data, 2)

# # A tibble: 2 x 7
# # Groups:   Sp [1]
#   Sp      L fit_eta se_eta fit_pi lwr_pi upr_pi
#   <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
# 1 Ed    16.3  -1.45  0.374  0.190  0.100  0.332
# 2 Ed    16.5  -1.47  0.368  0.187  0.0990 0.324
```

Визуализация в шкале логитов

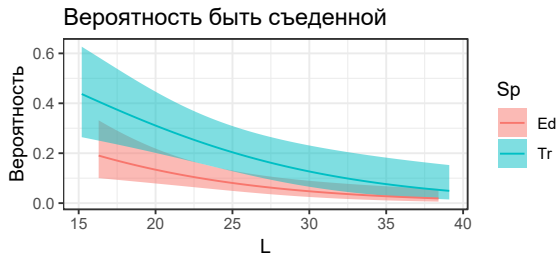
```
ggplot(new_data, aes(x = L, y = fit_eta, fill = Sp)) +  
  geom_ribbon(aes(ymin = fit_eta - 2 * se_eta, ymax = fit_eta + 2 * se_eta), alpha = 0.5) +  
  geom_line(aes(color = Sp))
```



Визуализация проведена для логитов, поэтому зависимость линейная, но по оси OY отложены значения от $-\infty$ до $+\infty$.

Визуализация в шкале вероятностей интуитивно понятнее

```
ggplot(new_data, aes(x = L, y = fit_pi, fill = Sp)) +  
  geom_ribbon(aes(ymin = lwr_pi, ymax = upr_pi), alpha = 0.5) +  
  geom_line(aes(color = Sp)) +  
  labs(y='Вероятность', title = 'Вероятность быть съеденной')
```

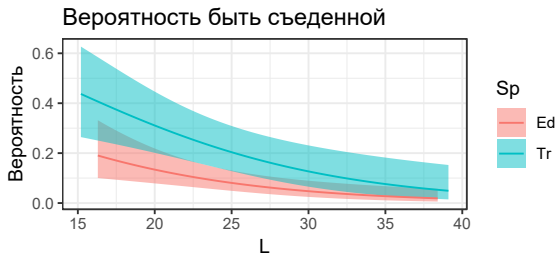


О чем говорит модель

Чем больше размер мидии, тем меньше вероятность быть съеденной.

Линия, соответствующая Tr , лежит выше линии Ed . Вероятность быть атакованной у Tr выше.

Значит звезды различают два вида мидий и размер жертвы для них имеет значение.



Что мы знаем о бинарных переменных

- ▶ Бинарные переменные-отклики могут обозначаться как угодно (+ или -; Да или Нет).
- ▶ Удобно кодировать бинарные переменные числами: 1 (событие произошло) или 0 (событие не произошло).
- ▶ Вместо бинарных обозначений в анализе используются непрерывные оценки вероятности.
- ▶ Вероятности можно перевести в отношения шансов.
- ▶ Отношения шансов заменяются логитами.

Что мы знаем о GLM с бинарной переменной-откликом

- ▶ GLM с бинарной переменной-откликом называют логистической регрессией.
- ▶ Параметры логистической регрессии подбираются методом максимального правдоподобия.
- ▶ Угловые коэффициенты логистической регрессии говорят о том, во сколько раз изменяется соотношение шансов для события при увеличении предиктора на единицу (или при переходе от базового уровня фактора к данному уровню).
- ▶ Оценить статистическую значимость модели можно с помощью анализа девиансы.
- ▶ Для визуализации результатов лучше проводить обратное логит-преобразование и изображать логистические кривые.

Что почитать

- ▶ Кабаков Р.И. R в действии. Анализ и визуализация данных на языке R. М.: ДМК Пресс, 2014.
- ▶ Quinn G.P., Keough M.J. (2002) Experimental design and data analysis for biologists, pp. 92-98, 111-130
- ▶ Zuur, A.F. et al. 2009. Mixed effects models and extensions in ecology with R. - Statistics for biology and health. Springer, New York, NY.