Регрессионный анализ для бинарных данных

Линейные модели…

Вадим Хайтов, Марина Варфоломеева

Мы рассмотрим

- Регрессионный анализ для бинарных зависимых переменных

Вы сможете

- Построить логистическую регрессионную модель, подобранную методом максимального правдоподобия
- Дать трактовку параметрам логистической регрессионной модели
- Провести анализ девиансы, основанный на логистической регрессии

Бинарные данные - очень распространенный тип зависимых переменных

- Вид есть - вида нет
- Кто-то в результате эксперимента выжил или умер
- Пойманное животное заражено паразитами или здорово
- Команда выиграла или проиграла

и т.д.

На каком острове лучше искать ящериц?

```
liz <- read.csv("data/polis.csv")
head(liz)
```

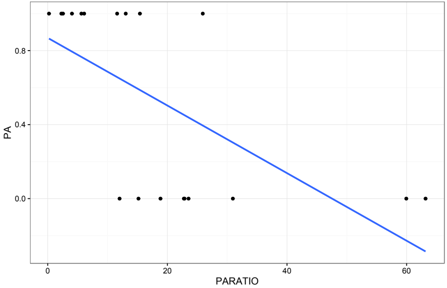
```
#      X.ISLAND PARATIO UTA PA PREDICT
# 1        Bota   15.41   P  1   0.555
# 2      Cabeza    5.63   P  1   0.915
# 3     Cerraja   25.92   P  1   0.111
# 4 Coronadito   15.17   A  0   0.568
# 5      Flecha   13.04   P  1   0.678
# 6    Gemelose   18.85   A  0   0.370
```

Пример взят из книги Quinn & Keugh (2002), Оригинальная работа Polis et al. (1998)

Зависит ли встречаемость ящериц от размера острова?

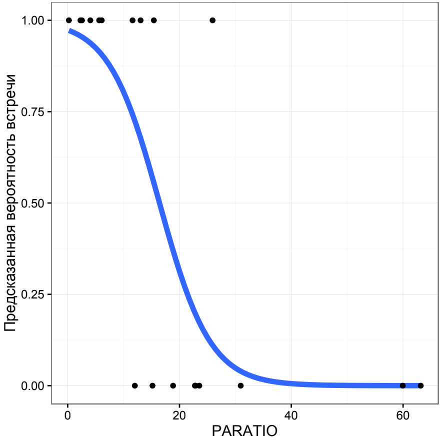
Обычную линейную регрессию подобрать можно,
Зависимая переменная:
PA - (есть ящерицы ”1” - нет ящериц ”0”)
Предиктор:
PARATIO (отношение периметра к площади)

```
fit <- lm(PA ~ PARATIO, data = liz)
summary(fit)
```



но она категорически не годится

Эти данные лучше описывает логистическая кривая



Логистическая кривая описывается такой формулой

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

Зависимую величину можно преобразовать в более удобную для моделирования форму

1. Дискретный результат: 1 или 0

Зависимую величину можно преобразовать в более удобную для моделирования форму

1. Дискретный результат: 1 или 0
2. Дискретные данные можно преобразовать в форму оценки вероятности события: $\pi = \frac{N_i}{N_{total}}$, непрерывная аеличина, варьирующая от 0 до 1

Зависимую величину можно преобразовать в более удобную для моделирования форму

1. Дискретный результат: 1 или 0
2. Дискретные данные можно преобразовать в форму оценки вероятности события: $\pi = \frac{N_i}{N_{total}}$, непрерывная аеличина, варьирующая от 0 до 1
3. Вероятность события можно выразить в форме шансов (odds): $odds = \frac{\pi}{1-\pi}$ варьируют от 0 до $+\infty$. *NB: Если шансы > 1, то вероятность события, что $y_i = 1$ выше, чем вероятность события $y_i = 0$. Если шансы < 1, то наоборот*. В обыденной речи мы часто использем фразы, наподобие такой "*шансы на победу 1 к 3*"

Зависимую величину можно преобразовать в более удобную для моделирования форму

1. Дискретный результат: 1 или 0
2. Дискретные данные можно преобразовать в форму оценки вероятности события: $\pi = \frac{N_i}{N_{total}}$, непрерывная аеличина, варьирующая от 0 до 1
3. Вероятность события можно выразить в форме шансов (odds): $odds = \frac{\pi}{1-\pi}$ варьируют от 0 до $+\infty$. *NB: Если шансы > 1, то вероятность события, что* $y_i = 1$ *выше, чем вероятность события* $y_i = 0$. *Если шансы < 1, то наоборот*. В обыденной речи мы часто использем фразы, наподобие такой "*шансы на победу 1 к 3*"
4. Шансы преобразуются в *Логиты* (logit): $ln\big(odds\big) = \ln\big(\frac{\pi}{1-\pi}\big)$ варьируют от $-\infty$ до $+\infty$. Логиты гораздо удобнее для построения моделей.

Что станет с логистической моделью после логит-преобразования?

Немного алгебры

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

Тогда

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

Что станет с логистической моделью после логит-преобразования?

Немного алгебры

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

Тогда

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

$$g(x) = \ln(\frac{e^z}{1 + e^z}) - \ln(1 - \frac{e^z}{1 + e^z})$$

Что станет с логистической моделью после логит-преобразования?

Немного алгебры

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

Тогда

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

$$g(x) = \ln(\frac{e^z}{1 + e^z}) - \ln(1 - \frac{e^z}{1 + e^z})$$

$$g(x) = \ln(\frac{e^z}{1 + e^z}) - \ln(\frac{1 + e^z - e^z}{1 + e^z}) = \ln(\frac{e^z}{1 + e^z}) - \ln(\frac{1}{1 + e^z})$$

Что станет с логистической моделью после логит-преобразования?

Немного алгебры

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

Тогда

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

$$g(x) = \ln(\frac{e^z}{1 + e^z}) - \ln(1 - \frac{e^z}{1 + e^z})$$

$$g(x) = \ln(\frac{e^z}{1 + e^z}) - \ln(\frac{1 + e^z - e^z}{1 + e^z}) = \ln(\frac{e^z}{1 + e^z}) - \ln(\frac{1}{1 + e^z})$$

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

Что станет с логистической моделью после логит-преобразования?

Немного алгебры

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

Тогда

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

$$g(x) = \ln(\frac{e^z}{1 + e^z}) - \ln(1 - \frac{e^z}{1 + e^z})$$

$$g(x) = \ln(\frac{e^z}{1 + e^z}) - \ln(\frac{1 + e^z - e^z}{1 + e^z}) = \ln(\frac{e^z}{1 + e^z}) - \ln(\frac{1}{1 + e^z})$$

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

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

Логистическая модель после логит-преобразования становится линейной

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

Остается только подобрать параметры этой линейной модели: β_0 (интерсепт) и β_1 (угловой коэффициент)

Метод максимального правдоподобия

Вспомним

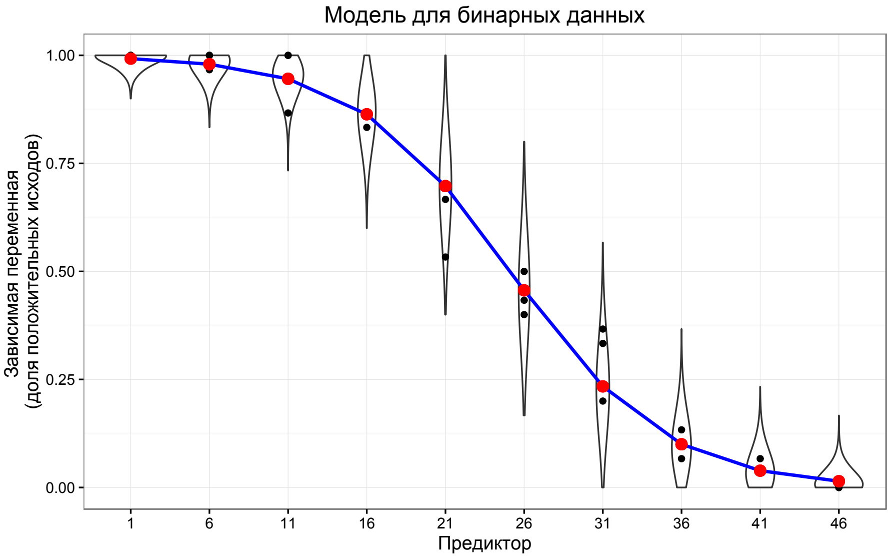
Если остатки не подчиняется нормальному распределению, то метод наименьших квадратов не работает.
В этом случае применяют *Метод максимального правдоподбия*

В результате итеративных процедур происходит подбор таких значений коэффициентов, при которых правдоподобие - вероятность получения имеющегося у нас набора данных - оказывается максимальным, при условии справедливости данной модели.

$$Lik(x_1, ..., x_n) = \Pi_{i=1}^{n} f(x_i; \theta)$$

где $f(x; \theta)$ - функция плотности вероятности с параметрами θ

Правдоподобие для биномиального распределения



Функция правдоподобия для биномиального распределения

Для случая биномиального распределения $x \in Bin(n, \pi)$ функция правдоподобия имеет следующий вид:

$$Lik(\pi|x) = \frac{n!}{(n-x)!x!}\pi^x(1-\pi)^{n-x}$$

отбросив константу, получаем:

$$Lik(\pi|x) \propto \pi^x(1-\pi)^{n-x}$$

Логарифм правдоподобия

Удобнее работать с логарифмом функции правдоподобия - *logLik* - его легче максимизировать. В случае биномиального распределения он выглядит так:

$$logLik(\pi|x) = xlog(\pi) + (n-x)log(1-\pi)$$

Подберем модель

```r
liz_model <- glm(PA ~ PARATIO , family="binomial", data = liz)
summary(liz_model)
```

```
#
# Call:
# glm(formula = PA ~ PARATIO, family = "binomial", data = liz)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -1.607   -0.638   0.237    0.433    2.099
#
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)     3.606      1.695    2.13    0.033 *
# PARATIO        -0.220      0.101   -2.18    0.029 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
#     Null deviance: 26.287  on 18  degrees of freedom
# Residual deviance: 14.221  on 17  degrees of freedom
# AIC: 18.22
#
```

summary() для модели, подобранной методом максимального правдоподобия

```
#
# Call:
# glm(formula = PA ~ PARATIO, family = "binomial", data = liz)
#
# Deviance Residuals:
#    Min      1Q  Median      3Q     Max
# -1.607  -0.638   0.237   0.433   2.099
#
# Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
# (Intercept)    3.606      1.695    2.13    0.033 *
# PARATIO       -0.220      0.101   -2.18    0.029 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
#     Null deviance: 26.287  on 18  degrees of freedom
# Residual deviance: 14.221  on 17  degrees of freedom
# AIC: 18.22
#
# Number of Fisher Scoring iterations: 6
```

“z value”” и “Pr(>z)”

z - это величина критерия Вальда (*Wald statistic*) - аналог t-критерия

Используется для проверки $H_0 : \beta_1 = 0$

$$z = \frac{\beta_1}{SE_{\beta_1}}$$

Сравнивают со стандартным нормальным распределением (z-распределение)

Дает надежные оценки p-value при больших выборках

Null deviance и Residual deviance

Имеющиеся данные позволяют “вписать” три типа моделей

“Насыщенная” модель - модель, подразумевающая, что каждая из n точек имеет свой собственный параметр, следовательно надо подобрать n параметров. Вероятность существования данных для такой модели равна 1.

$$logLik_{satur} = 0$$

$$df_{saturated} = n - npar_{saturated} = n - n = 0$$

“Нулевая” модель - модель, подразумевающая, что для описания всех точек надо подобрать только 1 параметр. $g(x) = \beta_0$.

$$logLik_{nul} \neq 0$$

$$df_{null} = n - npar_{null} = n - 1$$

“Предложенная” модель - модель, подобранная в нашем анализе $g(x) = \beta_0 + \beta_1 x$

$$logLik_{prop} \neq 0$$

$$df_{proposed} = n - npar_{proposed}$$

Null deviance и Residual deviance

Девианса - это оценка отклонения логарифма максимального правдоподобия одной модели от логарифма максимального правдоподобия другой модели

Остаточная девианса:
$Dev_{resid} = 2\left(logLik_{satur} - logLik_{prop}\right) = -2logLik_{prop}$

Нулевая девианса:
$Dev_{nul} = 2\left(logLik_{satur} - logLik_{nul}\right) = -2logLik_{nul}$

Проверим, совпадут ли со значениями из `summary()`

```
(Dev_resid <- -2*as.numeric(logLik(liz_model))) #Остаточная девианса
```

```
# [1] 14.2
```

```
(Dev_nul <- -2*as.numeric(logLik(update(liz_model, ~-PARATIO)))) #Нулевая дев
```

```
# [1] 26.3
```

Анализ девиансы

По соотношению нулевой девиансы и остаточной девиансы можно понять насколько статистически значима модель

В основе анализа девиансы лежит критерий G^2

$$G^2 = -2(logLik_{nul} - logLik_{prop})$$

```
(G2 <- Dev_nul - Dev_resid)
```

```
# [1] 12.1
```

Вспомним тест отношения правдоподобий:

$$LRT = 2ln(Lik_1/Lik_2) = 2(logLik_1 - logLlik_2)$$

> *Тест G^2 - это частный случай теста отношения правдоподобий (Likelihood Ratio Test)*

Свойства критерия G^2

- G^2 - это девианса полной (предложенной) и редуцированной модели (нулевой)

Свойства критерия G^2

- G^2 - это девианса полной (предложенной) и редуцированной модели (нулевой)
- G^2 - аналог частного F критерия в обычном регрессионном анализе

Свойства критерия G^2

- G^2 - это девианса полной (предложенной) и редуцированной модели (нулевой)
- G^2 - аналог частного F критерия в обычном регрессионном анализе
- G^2 - подчиняется χ^2 распределению (с параметом df = 1) если нулевая модель и предложенная модель не отличаются друг от друга.

Свойства критерия G^2

- G^2 - это девианса полной (предложенной) и редуцированной модели (нулевой)
- G^2 - аналог частного F критерия в обычном регрессионном анализе
- G^2 - подчиняется χ^2 распределению (с параметом df = 1) если нулевая модель и предложенная модель не отличаются друг от друга.
- G^2 можно использовать для проверки гипотезы о равенстве нулевой и остаточной девианс.

Задание

1. Вычислите вручную значение критерия G^2 для модели, описывающей встречаемость ящериц (`liz_model`)
2. Оцените уровень значимости для него

$$G^2 = -2(logLik_{nul} - logLik_{prop})$$

Решение

```
#Остаточная девианса
Dev_resid <- -2*as.numeric(logLik(liz_model))

#Нулевая девианса
Dev_nul <- -2*as.numeric(logLik(update(liz_model, ~-PARATIO)))

# Значение критерия
(G2 <- Dev_nul - Dev_resid)
```

```
# [1] 12.1
```

```
(p_value <- 1 - pchisq(G2, df = 1))
```

```
# [1] 0.000513
```

Решение с помощью функции anova()

```
anova(liz_model, test="Chi")
```

```
# Analysis of Deviance Table
#
# Model: binomial, link: logit
#
# Response: PA
#
# Terms added sequentially (first to last)
#
#
#         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
# NULL                       18       26.3
# PARATIO  1     12.1        17       14.2  0.00051 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Интерпретация коэффициентов логистической регрессии

Как трактовать коэффициенты подобранной модели?

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

```
coef(liz_model)
```

```
# (Intercept)     PARATIO
#        3.61       -0.22
```

β_0 - не имеет особого смысла, просто поправочный коэффициент

β_1 - *на сколько* единиц изменяется логарифм величины шансов (odds), если значение предиктора изменяется на единицу

Трактовать такую величину неудобно и трудно

Немного алгебры

посмотрим как изменится $g(x) = \ln(\frac{\pi(x)}{1-\pi(x)})$ при изменении предиктора на 1

$$g(x+1) - g(x) = ln(odds_{x+1}) - ln(odds_x) = ln(\frac{odds_{x+1}}{odds_x})$$

Задание: завершите алгебраическое преобразование

Решение

$$ln(\frac{odds_{x+1}}{odds_x}) = \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x = \beta_1$$

$$ln(\frac{odds_{x+1}}{odds_x}) = \beta_1$$

$$\frac{odds_{x+1}}{odds_x} = e^{\beta_1}$$

Полученная величина имеет определенный смысл

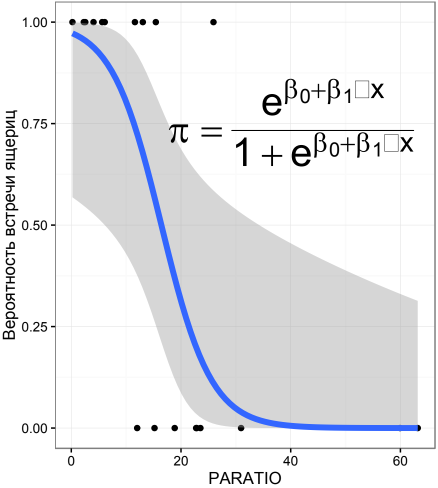
```
exp(coef(liz_model)[2])
```

```
# PARATIO
#   0.803
```

Во сколько раз изменяются шансы встретить ящерицу при увеличении отношения периметра острова к его площади на одну единицу. *NB: Отношение периметра к площади тем больше, чем меньше остров*.

Шансы изменяются в 0.803 раза. То есть, чем больше отношение периметра к площади, тем меньше шансов встретить ящерицу. Значит, чем больше остров, тем больше шансов встретить ящерицу

Подобранные коэффициенты позволяют построить логистическую кривую



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

Серая область - доверительный интервал для логистической регрессии
Доверительные интервалы для коэффициентов:

```
confint(liz_model) # для логитов
```

```
#                 2.5 %  97.5 %
# (Intercept)  1.006  8.0421
# PARATIO     -0.485 -0.0665
```

```
exp(confint(liz_model)) # для отношения
```

```
#                2.5 %   97.5 %
# (Intercept) 2.734 3109.275
# PARATIO     0.616    0.936
```

Задание:

Постройте график логистической регрессии для модели `liz_model` без использования `geom_smooth()`

Hint 1: Используйте функцию `predict()`, изучите значения параметра "type"

Hint 2: Для вызова справки напишите `predict.glm()`

Hint 3: Создайте датафрейм MyData с переменной `PARATIO`, изменяющейся от минимального до максимального значения `PARATIO`