МИРЭА
Российский технологический университет

Институт технологий управления

Программные средства анализа данных

01.03.05 Статистика

Профиль «Анализ данных в бизнесе и экономике»

Квалификация «бакалавр»

Бурцева Татьяна Александровна

Профессор кафедры статистики и математических методов в управлении

Burceva_t@mirea.ru

Москва, 2022

МИРЭА
Российский технологический университет

Тема 6. Логистическая регрессия в R

План лекции

1. Понятие логистической регрессии.
2. Пример реализации логистической регрессиия в R.
3. Другие виды регрессии в R.

МИРЭА
Российский технологический университет

1. Понятие логистической регрессии

Логистическая регрессия (Logistic regression)

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots \cdot \beta_k x_{k,i})}{(1+\exp(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots \cdot \beta_k x_{k,i}))}$$

- **Логистическая регрессия** или **логит-модель** — это статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём его сравнения с логистической кривой. То есть зависимая переменная номинативная (фактор с 2 градациями).
- Основываясь на независимых данных, такая модель определяет вероятность получения положительного ответа для реализации рискового события.
- Логистическая регрессия полезна для предсказания значений бинарной зависимой переменной по набору непрерывных и/или категориальных зависимых переменных.
- Логистическая регрессия - чрезвычайно эффективный механизм для вычисления вероятностей.

ОТЛИЧИЯ

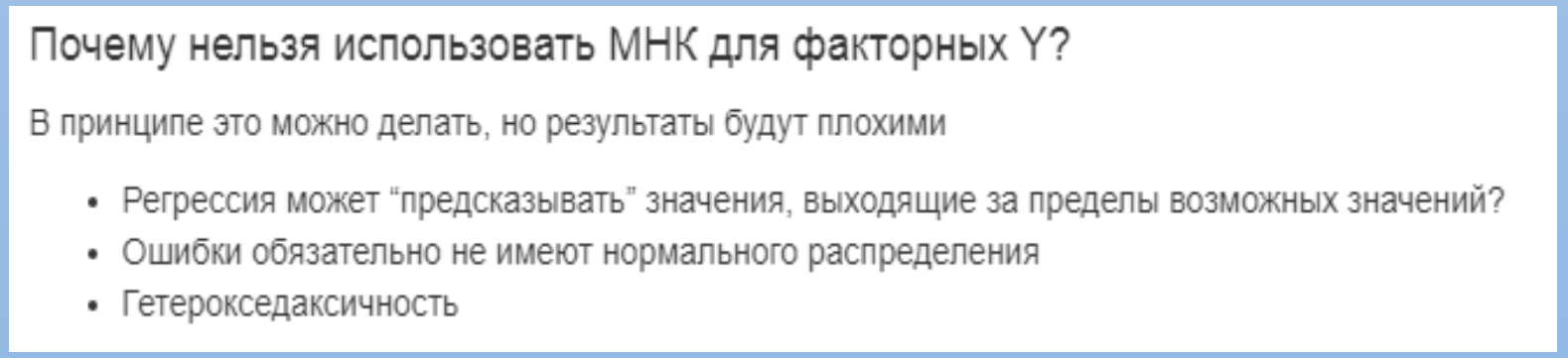
Логистическая регрессия - один из видов обобщенных линейных моделей (GLM). Допущение - ошибки (и Y) имеют биномиальное распределение.

Логистической регрессии от линейной является то, что в логистической регрессии ответ является конкретной константой. Тогда как в линейной регрессии ответ предсказания имеет непрерывный вид.

Зависимую переменную в логистической регрессии традиционно кодируют как 0 – 1, где 0 обозначает отсутствие какой-то характеристики, а 1 – ее наличие.

Логистическая регрессия в расчетах использует метод максимального правдоподобия (maximum likelihood estimation), тогда как линейная регрессия использует метод наименьших квадратов.

.



Логистическая регрессия может быть трех видов:

Бинарная логистическая регрессия.

Многомерная логистическая регрессия.

Порядковая логистическая регрессия.

Бинарная логистическая модель - только два признака, многомерная регрессия - несколько признаков.

Смысл логистической регрессии как модели

При логистической регрессии в качестве моделируемых значений зависимой переменной используется логарифм отношения шансов (odds) того, что $Y = 1$[2]. Регрессионный коэффициент – это изменение логарифма отношения шансов данной зависимой переменной на единицу изменения независимой переменной при постоянных значениях всех остальных независимых переменных.

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 * X_1)}}$$

Интепретация логистической регрессии - шансы (odds)

Шансы события - отношение вероятности выполнения события к вероятности его не выполнения. К примеру, шансы забеременеть - вероятность забеременеть делить на вероятность не забеременеть.

$$odds = \frac{P(event)}{P(noevent)}$$

Проблемы, возникающие в расчетах логистической регрессии

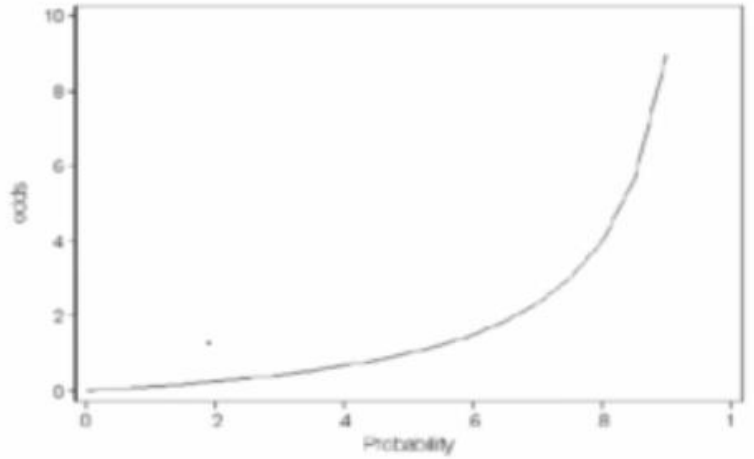
- полное разделение (complete separation)
- неполная информация по объясняющим факторам

У- вероятность, изменяется [0;1], а Xi нет, поэтому
нормировка данных = вероятность положительного исхода/ на вероятность отрицательного исхода

- пример

p	odds
.01	.010101
.2	.25
.3	.42
.4	.67
.25	.34
.5	1
.6	1.5

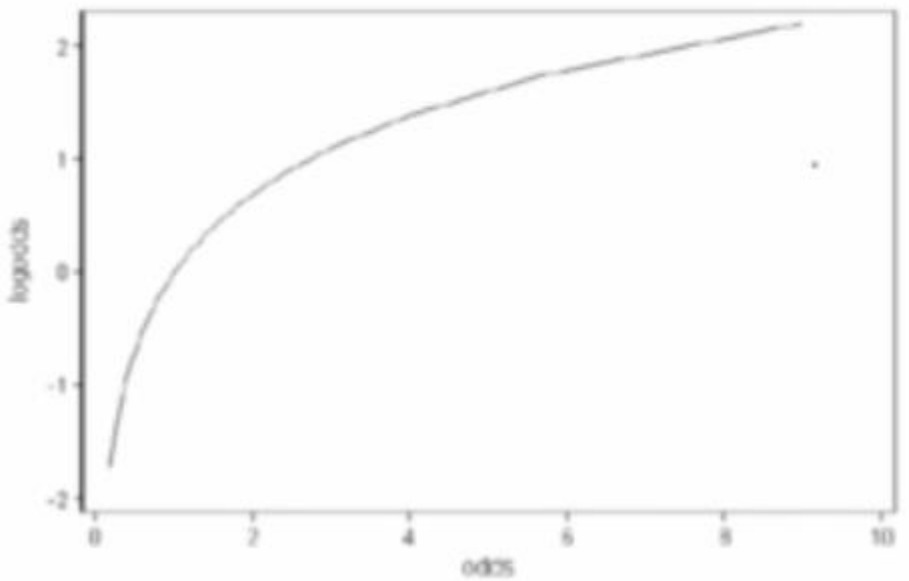
Odds– отношение вероятности успеха (Y = 1) к вероятности неудачи (Y = 0)
0.2 / (1 – 0.2) = 0.25



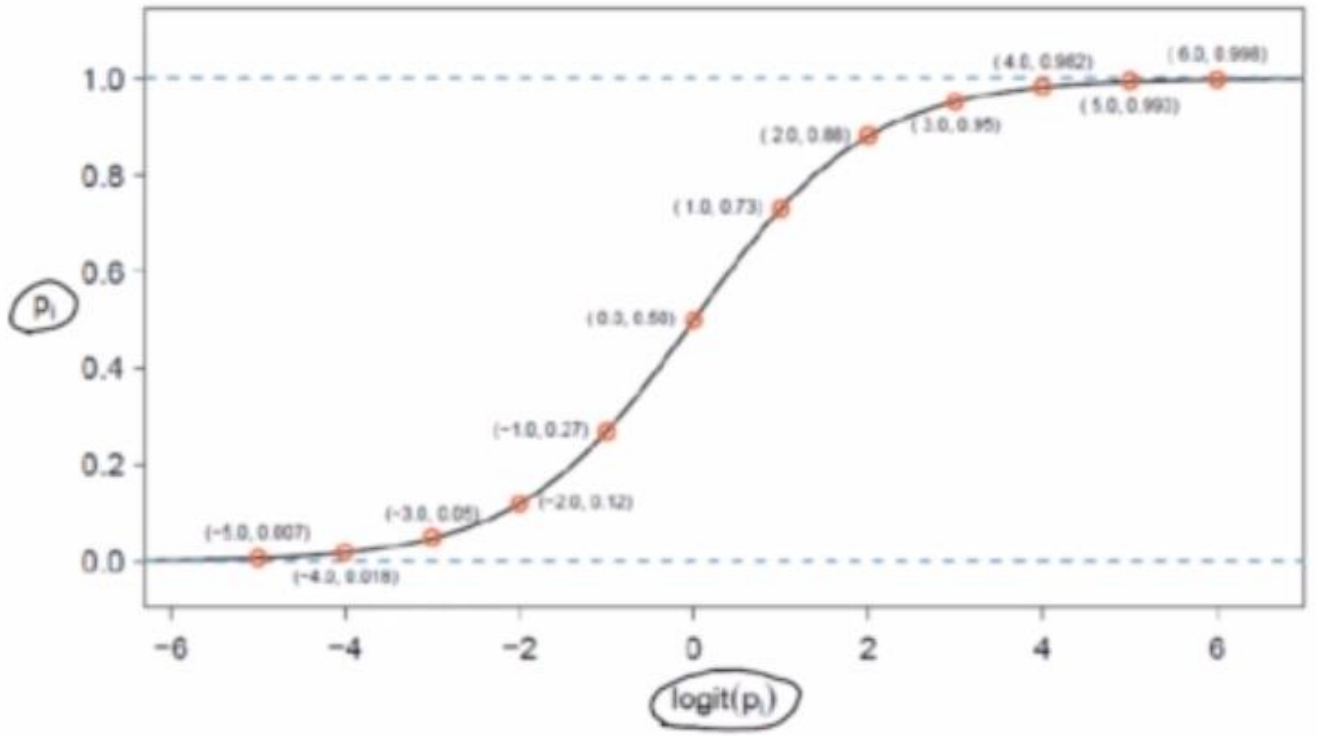
Логарифмируем

p	odds	logodds
.2	.25	-1.38
.3	.42	-.84
.4	.67	-.4
.5	1	0
.6	1.5	.4
.7	2.34	.84
.8	4	1.38

$$\log(0.2 / (1 - 0.2)) = \log(0.25) = -1.38$$



Зависимость p (вероятность) от Xi



$$\text{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_1 * \mathbf{x}_1$$

$$p = \exp(\beta_0 + \beta_1 * \mathbf{x}_1)/(1+\exp(\beta_0 + \beta_1 * \mathbf{x}_1)).$$

Требования к данным:

Независимость. Каждое наблюдение не должно зависеть от других наблюдений.

Линейность. Между хотя бы одной из непрерывных независимых переменных и логитом зависимой переменной должна наблюдаться линейная зависимость.

Отсутствие мультиколлинеарности. Независимые переменные не должны быть слишком связаны между собой.

Нет полного разделения. Значение одной переменной нельзя точно предсказать по значению другой переменной.

ИНСТРУМЕНТАРИЙ МОДЕЛИРОВАНИЯ

На данный момент для построения логистической регрессии используют языки программирования в сфере машинного и глубинного обучения R и Python.

МИРЭА
Российский технологический университет

2. Пример реализации логистической регрессиия в R

Пример

- Данные по 400 клиентам содержат информацию:
- У =1 (есть договор страхования)
- У=0 (нет договора страхования, так как не знают, будут или нет заключать договор)
- Х1 – размер страхового тарифа
- Х2 – страховая компания

Определим вероятность заключения страхового договора

- В переменную fit получим результаты моделирования
- fit <- glm(y ~ x1 + x2 + x1:x2, my_df, family = "binomial")

Результаты

-

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4210  -0.9050  -0.5809   1.0545   2.1412

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.36868    4.21385   0.325    0.745
x1          -2.34375    1.75545  -1.335    0.182
x2          -0.07427    1.21144  -0.061    0.951
x1:x2        0.45718    0.50172   0.911    0.362
```

МИРЭА
Российский технологический университет

Определим вероятность заключения договора о страховании

- #посчитали logit(p)
- my_df$prob<-1.36868-2.34375*my_df$x1-0.07427*my_df$gx2+0.45718*my_df$x1*my_df$x2
- #посчитали p
- my_df$p<-exp(my_df$prob)/(1+exp(my_df$prob))

Замечания:

- Поскольку независимые переменные не могут быть равны нулю, то свободный член в данном случае лишён смысла.
- Доверительные интервалы для коэффициентов вычисляются при помощи функции confint().
- Команда exp(confint(fit.reduced)) позволит рассчитать доверительные интервалы для всех коэффициентов модели для каждого коэффициента в единицах отношения шансов.

В R реализовано несколько дополнительных методов и разновидностей логистической регрессии:

- *устойчивая (robust) логистическая регрессия.* Функцию `glmRob()` из пакета `robust` можно использовать для подгонки устойчивых обобщенных регрессионных моделей, в том числе и устойчивой логистической регрессии. Этот метод может быть полезен при подборе логистических регрессионных моделей для данных с выбросами и влиятельными наблюдениями;
- *мультиномиальная (multinomial) логистическая регрессия.* Если зависимая переменная имеет больше двух неупорядоченных значений (например, женат/вдов/разведен), можно рассчитать мультиномиальную логистическую регрессию при помощи функции `mlogit()` из пакета `mlogit`;
- *порядковая (ordinal) логистическая регрессия.* Если зависимая переменная представляет собой упорядоченный фактор (например, платежеспособность по кредиту: слабая/хорошая/отличная), можно использовать порядковую логистическую регрессию (функция `lrm()` из пакета `rms`).

МИРЭА
Российский технологический университет

3. Другие виды регрессии в R

Обобщённые линейные модели и функция glm()

МИРЭА
Российский технологический университет

Обобщённые линейные модели и функция glm()

Допустим, вы хотите смоделировать связь между зависимой переменной Y и набором из p независимых переменных $X_1 \ldots X_p$. В стандартной линейной модели вы предполагали, что Y имеет нормальное распределение и тип связи описывается таким уравнением:

$$\mu_Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j .$$

Это значит, что условное среднее зависимой переменной представляет собой линейную комбинацию значений независимых переменных. Параметры, определяющие ожидаемое изменение переменной Y на единицу изменения X_j, обозначены как β_j, а β_0 – это ожидаемое значение переменной Y, когда все независимые переменные равны нулю. Вы можете предсказать среднее значение распределения значений Y для наблюдений с заданным набором значений X, сообщая определенные веса переменным X и суммируя их.

Обратите внимание, что мы не делали никаких предположений относительно распределения значений независимых переменных X_j. Они, в отличие от Y, не обязательно должны иметь нормальное распределение. На самом деле они часто бывают категориальными (например, при дисперсионном анализе). Кроме того, допускаются нелинейные комбинации независимых переменных. Нередко в уравнение входят такие независимые переменные как X^2 или $X_1 \times X_2$. Что здесь важно, так это линейность комбинаций параметров (β_0, β_1, ... β_p).

При создании обобщенных линейных моделей мы подбираем модели в виде

$$g(\mu_Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j ,$$

где $g(\mu_Y)$ – это функция условного среднего (называемая связующей функцией). Кроме того, вы отказываетесь от предположения о нормальном распределении значений Y. Вместо этого вы подразумеваете, что Y подчиняется распределению из семейства экспоненциальных распределений. Вы задаете связующую функцию и вероятностное распределение, а параметры оцениваются при помощи итеративного алгоритма максимального правдоподобия.

Функция glm()

Обобщенные линейные модели обычно подгоняются в R при помощи функции `glm()` (хотя доступны и другие функции). Формат применения этой функции сходен с таковым для функции `lm()`, но имеет дополнительные параметры. Основной формат функции таков:

```
glm(formula, family=family(link=function), data=)
```

Типы распределений вероятностей (*family*) и соответствующие связующие функции (*function*) по умолчанию приведены в табл.

Параметры функции `glm()`

Семейство распределений	Связующая функция по умолчанию
binomial	(link = "logit")
gaussian	(link = "identity")
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

Функция `glm()` позволяет подгонять разные распространенные модели, включая логистическую и пуассоновскую регрессии, а также модели для анализа выживания (здесь не рассматривается).

Логистическая регрессия подходит для дихотомических зависимых переменных (0, 1). Модель предполагает, что Y имеет биномиальное распределение и что можно подогнать линейную модель следующего вида:

$$\log_e\left(\frac{\pi}{1-\pi}\right)=\beta_0+\sum_{j=1}^{p}\beta_j X_j,$$

где $\pi=\mu_Y$ – это условное среднее Y (то есть вероятность того, что $Y=1$ при данном наборе значений X), $(\pi/1-\pi)$ – это отношение шансов того, что $Y=1$, а $\log(\pi/1-\pi)$ – это логарифм отношения шансов, или *логит*. В данном случае $\log(\pi/1-\pi)$ – это связующая функция, вероятностное распределение биномиальное, а логистическая регрессионная модель может быть подогнана при помощи команды

```
glm(Y~X1+X2+X3, family=binomial(link="logit"), data=mydata)
```

Обобщенные линейные модели и функция glm()

Пуассоновская регрессия применяется в случае счетной зависимой переменной. Пуассоновская регрессионная модель подразумевает, что Y имеет пуассоновское распределение и что можно подогнать линейную модель вида

$$\log_e(\lambda) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j ,$$

где λ – это среднее (и дисперсия Y). В данном случае связующая функция имеет вид $\log(\lambda)$, вероятностная функция пуассоновская, а пуассоновская регрессионная модель может быть подобрана при помощи команды

```
glm(Y~X1+X2+X3, family=poisson(link="log"), data=mydata)
```

В R реализовано несколько полезных дополнений, которые полезны, когда данные содержат выбросы и влиятельные наблюдения.

Для подбора устойчивой обобщённой линейной модели, при наличии выбросов и влиятельных наблюдений, можно использовать функцию glmRob() из пакета robust.

Функции, которые используются вместе с функцией glm()

МИРЭА Российский технологический университет

Функция	Описание
`summary()`	Выводит на экран подробные результаты для подогнанной модели
`coefficients(), coef()`	Выводит параметры модели (свободный член и регрессионные коэффициенты)
`confint()`	Доверительные интервалы для параметров модели (по умолчанию 95%)
`residuals()`	Выводит остатки подогнанной модели
`anova()`	Создает таблицу дисперсионного анализа для сравнения двух моделей
`plot()`	Создает диагностические диаграммы для оценки соответствия модели данным
`predict()`	Использует подогнанную модель для предсказания значений зависимой переменной для нового набора данных

Вопросы по теме

1. В чем отличия логистической регрессии от линейной?
2. Какие виды логистической регрессии бывают?
3. Как реализуется логистическая регрессия в R?
4. Какие команды реализуют логистическую регрессию в R?
5. Что такое пуассоновская регрессия?
6. Обобщённые линейные модели регрессии почему нужно использовать?

Литература

1. Эконометрика и эконометрическое моделирование в EXCEL и R: учебник/Л.О. Бабешко, И.В. Орлова. – Москва: ИНФРА-М, 2021. – 300 с.: ил. – (Высшее образование: Магистратура). –DOI 10.12737/1079837.
2. Роберт И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил. ISBN 978-5-947060-077-1