1 Модели логистической регрессии

1.1 Предварительные сведения по логистической регрессии

Рассмотрим теперь модели, в которых объясняемая переменная Y принимает только два значения, например, 0 или 1[1].

Ситуации такого рода возникают, при исследовании влияния тех или иных субъективных и объективных факторов на наличие или отсутствие некоторого признака у объекта исследования (наличие или отсутствие в семье автомобиля, занятый-безработный, фирма обанкротилась в течение некоторого периода или нет, наличие или отсутствие заболевания и т.д.).

Пусть имеется совокупность n наблюдений над каким-то объектом:

$$\begin{pmatrix} y_1 & x_{11} & \dots & x_{1p} \\ y_2 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ y_n & x_{n1} & \dots & x_{np} \end{pmatrix},$$

где Y – результирующая (зависимая) переменная, X_j – j-я независимая переменная (объясняющая переменная, фактор), $j = \overline{1, p}$.

Факт наличия или отсутствия какого-то признака в i-м наблюдении удобно индексировать числами 1 (наличие признака) и 0 (отсутствие признака), то есть $y_i = 0$ или 1.

Классическая линейная модель множественной регрессии:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = \overline{1, n},$$

для оценки влияния факторов на Y не подходит.

Причины этого:

1) При обычном предположении $M\left(\varepsilon_i / X_i^T\right) = 0, \quad i = \overline{1, n}$ получаем

[1]В общем случае зависимая переменная может принимать и другие значения. Если эти значения измерены в категориальной шкале (переменная принимает значения да/нет, сдал/провалился, живой/мертвый или плохой/хороший/превосходный, республиканец/демократ/независимый), то строится логистическая регрессия, если в порядковой (счетной) ((например, число дорожно-транспортных происшествий за неделю), то пуассоновская регрессия.

$$M\left(y_i / X_i^T\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} = X_i^T B, \quad i = \overline{1,n},$$

где $X_i^T = (1, x_{i1}, x_{i2}, ..., x_{ip})$, $B = \begin{pmatrix} \beta_0 \\ ... \\ \beta_p \end{pmatrix}$.

С другой стороны, y_i – случайная величина, принимающая значения 0 и 1. Следовательно, ее математическое ожидание равно:

$$M\left(y_i / X_i^T\right) = 1 \cdot P(y_i = 1 / X_i^T) + 0 \cdot P(y_i = 0 / X_i^T) = P(y_i = 1 / X_i^T)$$

Следовательно, $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} = P(y_i = 1 / x_i)$ и тогда должно выполняться

$$0 \le \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} \le 1$$

2) Проблема гетероскедастичности остатков

$$D\left(\varepsilon_i / X_i^T\right) = X_i^T B \cdot \left(1 - X_i^T B\right) \ne const.$$

3) Пусть y_i – показывает наличие или отсутствие заболевания у i-го пациента в возрасте x_i. Естественно предположить, что вероятность заболевания у пациента возрастает с возрастом. Если при этом использовать линейную модель

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = \overline{1,n},$$

то

$$M\left(y_i / x_i\right) = P(y_i = 1 / x_i) = \beta_0 + \beta_1 x_i,$$

поэтому, если значение x_i увеличить на единицу, то вероятность наличия заболевания увеличится на величину равную

$$\beta_0 + \beta_1 (x_i + 1) - (\beta_0 + \beta_1 x_i) = \beta_1$$

независимо от того, сколь большим или малым является возраст пациента x_i. Между тем такое положение вряд ли можно считать оправданным. Скорее можно предположить, что для молодых пациентов наличие заболевания – редкость, и некоторое увеличение возраста лишь ненамного увеличит вероятность приобретения заболевания. Для пациентов старшего возраста возрастание вероятности наличия заболевания также не может быть

существенным, поскольку такие пациенты, как правило, уже имеют заболевание. Большее влияние увеличения возраста на возрастание вероятности наличия заболевания должно наблюдаться для пациентов среднего возраста, т.е. в «переходной зоне».

Для описания таких зависимостей больше подходят функции, значения которых ограничены на интервале [0, 1], график которых имеет S-образную форму.

Примерами таких функций являются

$g(x) = \dfrac{e^x}{1+e^x} = \dfrac{1}{1+e^{-x}}$ – функция стандартного логистического распределения (логит-модель).

$\Phi(x) = \dfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{x^2}{2}} dx$ – функция стандартного нормального распределения $N(0,1)$ (пробит-модель).

$G(x) = 1 - e^{-e^x}$ – функция стандартного распределения экстремальных значений (минимума) I-го типа (распределение Гомпертца, гомпит-модель).

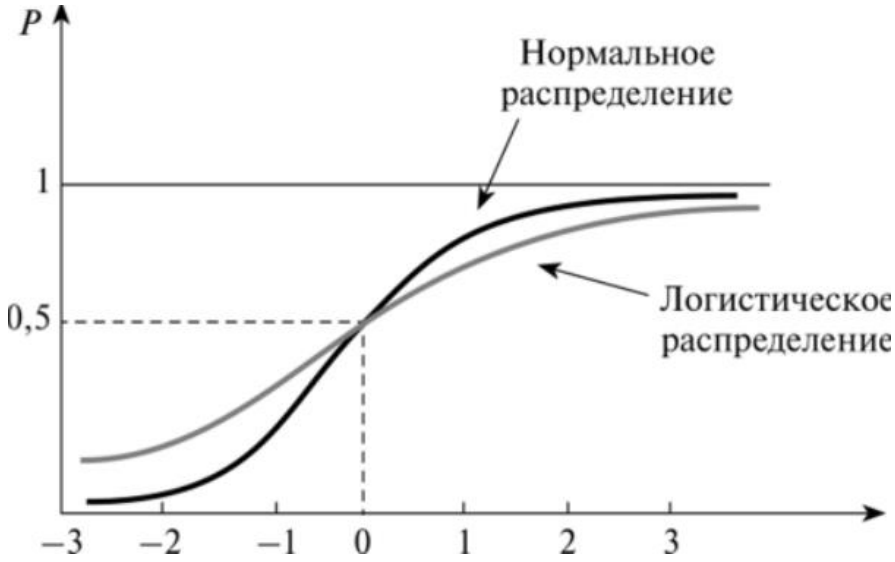


Рисунок 1 – Графики функций $g(x)$ и $\Phi(x)$

Модель логистической регрессии в случае использования **логит-модели** имеет вид

$$M(y_i / X_i^T) = P\left(y_i = 1/ X_i^T\right) = \frac{e^{X_i^T B}}{1+e^{X_i^T B}} \text{ или } \frac{1}{1+e^{-X_i^T B}}.$$

В случае одной независимой переменой модель примет вид

$$M(y_i / x_i) = P\left(y_i = 1/ x_i\right) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}.$$

Отсюда можно выразить линейную функцию (так называемый логит) для логарифма отношения шансов, что $y_i = 1$ к $y_i = 0$

$$\beta_0 + \beta_1 x_i = \ln\frac{P(y_i = 1/ x_i)}{1-P(y_i = 1/ x_i)} = \ln\frac{P(y_i = 1/ x_i)}{P(y_i = 0/ x_i)}.$$

1.2 Оценка неизвестных параметров β_0 и β_1 методом максимального правдоподобия

Отступление:

Метод максимального правдоподобия

Пусть имеются наблюдения за 4 дня за случайной величиной – числом телефонных звонков *X*, в первый день число звонков было 0, то есть x_1=0, во второй день x_2=1, в третий x_3=2 звонка, в четвертый x_4=0 звонков. Так же полагаем, что *X* имеет закон распределения:

X	0	1	2
p	*p*	2*p*	1-3*p*

Требуется найти оценку p.

Полагаем, что число звонков в разные дни независимы.

Составим функцию правдоподобия, которая позволяет найти такое значение оценки p, при котором вероятность получить наблюдаемые значения *X* максимальна:

Составим вероятность того, что

$$P(x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 0) = P(x_1 = 0)P(x_2 = 1)P(x_3 = 2)P(x_4 = 0) =$$
$$= p \cdot 2p \cdot (1-3p) \cdot p = 2p^3(1-3p) = L(p) \to \max.$$

Удобнее работать с логарифмом

$$\ln L(p) = \ln 2 p^3 (1-3p) = \ln 2 + 3\ln p + \ln(1-3p) \to \max.$$

Находим производную

$$\left(\ln L(p)\right)' = \left(\ln 2 + 3\ln p + \ln(1-3p)\right)' = \frac{3}{p} - \frac{3}{1-3p} = \frac{3-9p-3p}{p(1-3p)} =$$

$$= \frac{3-12p}{p(1-3p)} = 0.$$

Откуда $p = 1/4.$

Теперь вернемся к нашей задаче, которая состоит в оценивании параметров модели логистической регрессии:

$$y_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} + \varepsilon_i, \quad i = \overline{1,n}.$$

В дальнейшем, поскольку значения параметров β_0 и β_1 оцениваются по выборкам из *n* наблюдений, то оценки параметров будем обозначать как b_0 и b_1. Предполагаем, что при фиксированных значениях объясняющих переменных в *n* наблюдениях, случайные ошибки $\varepsilon_1,...,\varepsilon_n$ статистически независимы и $M(\varepsilon_i / x_i) = 0$, так что

$$P(y_i = 1/x_i) = M(y_i / x_i) = \frac{e^{b_0+b_1 x_i}}{1+e^{b_0+b_1 x_i}}.$$

Следовательно, статистически независимы и $y_1,...,y_n$. Поэтому условная совместная вероятность получения конкретного набора наблюдений $y_1,...,y_n$ (конкретного набора нулей и единиц) равна произведению

$$P(y_1 = 1(0)/x_1, y_2 = 0(1)/x_2, y_3 = 1(0)/x_3,..., y_n = 0(1)/x_n) =$$
$$= P(y_1 = 1(0)/x_1)P(y_2 = 0(1)/x_2)P(y_3 = 1(0)/x_3)\cdot...\cdot P(y_n = 0(1)/x_n) =$$
$$= \prod_{i=1}^{n} \left[P(y_i = 1/x_i)\right]^{y_i} \left[P(y_i = 0/x_i)\right]^{1-y_i} =$$
$$= \prod_{i=1}^{n} \left[P(y_i = 1/x_i)\right]^{y_i} \left[1 - P(y_i = 1/x_i)\right]^{1-y_i} =$$

$$= \prod_{i=1}^{n} \left[\frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}} \right]^{y_i} \left[1 - \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}} \right]^{1 - y_i} =$$

$$= \prod_{i=1}^{n} \left[\frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}} \right]^{y_i} \left[\frac{1}{1 + e^{b_0 + b_1 x_i}} \right]^{1 - y_i} = L(B).$$

Правая часть этого выражения является функцией от вектора $B = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$ и интерпретируется как функция правдоподобия $L(B)$ параметров b_0, b_1.

При различных наборах значений b_0, b_1 получаются различные $L(B)$, то есть при фиксированных x_i, вероятность наблюдать конкретный набор значений $y_1,...,y_n$ может быть более высокой или более низкой, в зависимости от значения b_0 и b_1. Метод максимального правдоподобия предлагает в качестве оценки вектора параметров B использовать значения b_0 и b_1, максимизирующие функцию правдоподобия, так что

$$L(\hat{B}) = \max_B L(B) = \max_B \prod_{i=1}^{n} \left[\frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}} \right]^{y_i} \left[\frac{1}{1 + e^{b_0 + b_1 x_i}} \right]^{1 - y_i}.$$

Использование свойства монотонного возрастания функции $ln(x)$ позволяет найти то же самое значение $\hat{B} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \end{pmatrix}$, максимизируя логарифмическую функцию правдоподобия $ln\ L(B)$. В нашем случае

$$\ln L(B) = \sum_{i=1}^{n} y_i \ln \left[\frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}} \right] + \sum_{i=1}^{n} (1 - y_i) \ln \left[\frac{1}{1 + e^{b_0 + b_1 x_i}} \right].$$

Преобразуем функцию к виду:

$$\ln L(B) = \sum_{i=1}^{n} y_i \ln\left[\frac{e^{b_0+b_1x_i}}{1+e^{b_0+b_1x_i}}\right] + \sum_{i=1}^{n}(1-y_i)\ln\left[\frac{1}{1+e^{b_0+b_1x_i}}\right] =$$

$$= \sum_{i=1}^{n} y_i(b_0+b_1x_i) - y_i \ln\left[1+e^{b_0+b_1x_i}\right] - \sum_{i=1}^{n}(1-y_i)\ln\left[1+e^{b_0+b_1x_i}\right] =$$

$$= \sum_{i=1}^{n} y_i(b_0+b_1x_i) - y_i \ln\left[1+e^{b_0+b_1x_i}\right] - \ln\left[1+e^{b_0+b_1x_i}\right] + y_i \ln\left[1+e^{b_0+b_1x_i}\right] =$$

$$= \sum_{i=1}^{n} y_i(b_0+b_1x_i) - \ln\left[1+e^{b_0+b_1x_i}\right].$$

1.3 Построение и анализ модели логистической регрессии на примере

Далее процесс вычисления продолжим в R для примера, исходные данные которого приведены в таблице 1.

Таблица 1 – Исходные данные

Номер пациента, *i*	Возраст пациента, *x*	Наличие заболевания, *y*
1	25	0
2	29	0
3	30	0
4	31	0
5	32	0
6	41	0
7	41	0
8	42	0
9	44	1
10	49	1
11	50	0
12	59	1
13	60	0
14	62	0
15	68	1
16	72	0
17	79	1
18	80	0
19	81	1
20	84	1

Перед тем, как строить модель логистической регрессии в R-studio, можно набрать таблицу 1 в Excel, для удобства переименовать переменную **Возраст** в **age**, а переменную **Наличие заболевания** в **disease**. Затем необходимо сохранить файл с исходной таблицей как .csv файл. Для этого в Excel нужно сохранить файл как .csv (разделитель – запятые). Пусть, например, этот файл сохранен под именем logit.csv в папке R/examples. Затем открываем этот файл в R-studio, набрав соответствующие команды

```
logit <- read.csv("~/R/examples/logit.csv", sep=";")
View(logit).
```

Теперь воспользуемся функцией glm, написав и выполнив код

```
model1 <- glm(data = logit, disease ~ age,
                    family = binomial(link = "logit") )
summary(model1)
```

Результаты показаны на рисунке 2.

```
Call:
glm(formula = disease ~ age, family = binomial(link = "logit"),
    data = logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6136  -0.6591  -0.4310   0.7856   1.8118

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.37210    1.96555  -2.224   0.0261 *
age          0.06696    0.03223   2.077   0.0378 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.898  on 19  degrees of freedom
Residual deviance: 20.201  on 18  degrees of freedom
AIC: 24.201

Number of Fisher Scoring iterations: 4

> |
```

Рисунок 2 – Результаты построения модели логистической регрессии

Интерпретация параметров модели

Из рисунка 2 видно, что значения параметров b_0=-4,372, b_1=0,067, то есть

$$P(y_i = 1/x_i) = \frac{e^{-4,372+0,067x_i}}{1+e^{-4,372+0,067x_i}}.$$

Применив функцию predict с параметром type="response"[1]:

```
predict(model1, newdata = data.frame(age=c(50, 80)), type="response")
```

получили, что пациент в возрасте, например, 50 лет имеет заболевание с вероятностью 26,4%, вероятность отсутствия заболевания 100%-26,4%=74,6%, а в возрасте 80 лет вероятность заболевания уже 72,8%, вероятность его отсутствия 27,2%:

Как уже отмечалось выше

$$b_0 + b_1 x_i = \ln\frac{P(y_i = 1/x_i)}{1-P(y_i = 1/x_i)} = \ln\frac{P(y_i = 1/x_i)}{P(y_i = 0/x_i)}$$

или проще

$$b_0 + b_1 x_i = \ln\frac{P}{1-P}.$$

Отношение $\frac{P}{1-P} = Odds$ называется **шансами** события $y_i = 1$. Например, если P=2/3, то Odds=(2/3)/(1/3)=2. То есть шансы за то, что $y_i = 1$ против того, что $y_i = 0$ равны 2 к 1 или в 2 раза выше. Логарифм от Odds называют логитом (logit(P)). Если Odds=1, то logit(P)=0, то есть шансы для события $y_i = 1$ равны 1 к 1. Если logit(P)>0, то больше шансов, что событие $y_i = 1$ произойдет, если logit(P)<0, то больше шансов, что это событие не произойдет.

Логит-модель линейна в отношении логита

$$\ln\frac{P}{1-P} = b_0 + b_1 x_i \text{ или } \ln\frac{P}{1-P} = -4,372 + 0,067x_i.$$

Отсюда вытекает, что изменение значения объясняющей переменной на единицу приводит к изменению значения $\ln\frac{P}{1-P}$ в среднем на b_1 единиц,

[1] Опция type = “response” сообщает R, что необходимо выводить вероятности в виде P(Y = 1|X), а не какую-то другую информацию вроде логита.

а Odds в e^{b_1} раз. Иначе говоря, при этом шансы за то, что $y_i = 1$, против того, что $y_i = 1$, возрастают приблизительно в e^{b_1} раз. В нашем случае можно найти экспоненты от коэффициентов модели с помощью команды

exp(coef(model1))

В результате

```
(Intercept)         age
 0.01262473  1.06924849
```

То есть при увеличении возраста на 1 год, шансы заболеть увеличиваются в среднем в 1,069 раз**.**

Оценка качества модели

Для оценки качества полученной модели используют понятие тривиальной (нулевой) модели (null model) и насыщенной модели (saturated model).

Тривиальная модель выглядит следующим образом

$$P(y_i = 1/x_i) = \frac{e^{b_0}}{1+e^{b_0}}. \qquad (1)$$

Логарифм функции правдоподобия для такой модели будет следующим

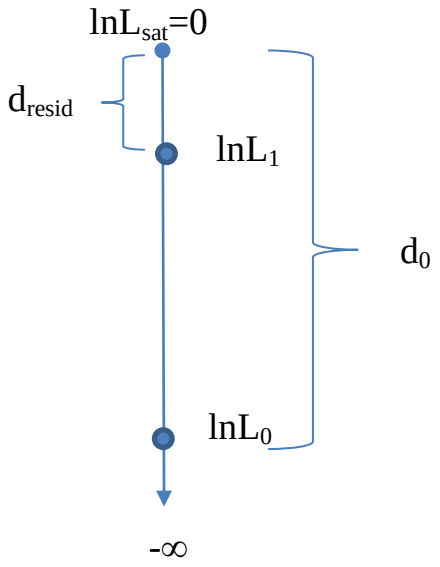
$$\ln L_0 = \sum_{i=1}^{n} \left(y_i b_0 - \ln\left[1+e^{b_0}\right]\right) = b_0 \sum_{i=1}^{n} y_i - n \ln\left[1+e^{b_0}\right]$$

Максимальное значение эта функция примет, при $b_0 = \ln \frac{\sum_{i=1}^{n} y_i}{n - \sum_{i=1}^{n} y_i}$.

Насыщенная модель была бы в состоянии генерировать (оцененные) вероятности, которые в точности соответствовали бы наблюдаемым значениям, тогда все вероятности в логарифме правдоподобия были бы равными единице, так что логарифм правдоподобия был бы в точности равен

нулю, то есть $\ln L_{sat} = 0$ – это максимальное значение логарифма функции правдоподобия.

Поскольку мы использовали для оценивания таких моделей метод максимального правдоподобия, то естественным представляется сравнение максимума логарифмической функции правдоподобия выбранной модели с аналогичными максимумами функций для тривиальной и насыщенной моделей. Тогда максимум логарифмической функции правдоподобия для нашей модели будет между $\ln L_0 < 0$ и $\ln L_{sat} = 0$. Пусть $\ln L_1$ – максимум логарифма функции правдоподобия для нашей модели. Чем больше разность между $\ln L_0$ и $\ln L_1$, тем больше расширенная модель дополняет очень ограниченную тривиальную модель.



Девиансой называется оценка разницы логарифмов правдоподобий.

Различают остаточную **девиансу**

$$d_{resid} = 2(\ln L_{sat} - \ln L_1) = 2(0 - \ln L_1) = -2\ln L_1$$

и нулевую девиансу

$$d_0 = 2(\ln L_{sat} - \ln L_0) = 2(0 - \ln L_0) = -2\ln L_0$$

Значения девианс выводится в отчете функции summary()

```
    Null deviance: 25.898  on 19  degrees of freedom
Residual deviance: 20.201  on 18  degrees of freedom
```

Сравнение нулевой и остаточной девианс позволяет судить о статистической значимости модели в целом.

Такое сравнение проводится с помощью теста отношения правдоподобий (likelihood ratio test, LRT). Нулевая гипотеза здесь состоит в том, что разница между нулевой и остаточной девиансами не существенна. Величина $d_0 - d_{resid} = -2\ln L_0 + 2\ln L_1 = 2(\ln L_1 - \ln L_0)$ распределена по χ^2 со степенью свободы $df=df_1-df_0=n-2-(n-1)=1$ (здесь 2 и 1 – число параметров в нашей и тривиальной модели). Если окажется, что $d_0 - d_{resid} > \chi^2(0.05; 1)$, то нулевая гипотеза отвергается на уровне значимости 0,05. Протестируем значимость модели в целом в R. Для этого необходимо создать тривиальную модель

```
model0<-glm(data=logit, disease~1, family=binomial(link="logit"))
```

и передать эту модель и тестируемую (model1) в функцию anova() с указанием параметра test = “Chi”. Получим

```
> anova(model0, model1, test="Chi")
Analysis of Deviance Table

Model 1: disease ~ 1
Model 2: disease ~ age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        19     25.898
2        18     20.201  1   5.6964    0.017 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Так как Pr(>|Chi|) = 0.017<0.05, то на уровне значимости 0,05 нулевая гипотеза отвергается и делаем вывод о том, что наша модель значительно лучше тривиальной.

Так же можно рассчитать следующие показатели для оценки качества модели

$$PseudoR^2 = 1 - \frac{1}{1 + 2(\ln L_1 - \ln L_0)/n},$$

$McFaddenR^2 = LRI = 1 - \frac{LnL_1}{LnL_0}$ – индекс отношения правдоподобия.

Оба этих показателя изменяются в пределах от 0 до 1. Чем лучше качество модели по сравнению с тривиальной, тем ближе показатель к 1.

Задание: в R рассчитать значения этих показателей.

Оценка значимости параметров модели

На рисунке 2 в столбце Pr(>|z|) для переменной age приведена вероятность, подтверждающая гипотезу о том, что для описания поведения переменной disease подходит модель (1), то есть age не оказывает значимого влияния на наличие или отсутствие заболевания. Так как Pr(>|z|)=0,0378<0,05, то можно говорить о том, что такую гипотезу надо отвергнуть и оценка параметра β_1 модели значимо отличается от 0 на уровне значимости 0,05.

Точность предсказания модели

Точность предсказания модели может быть найдена с помощью таблицы сопряженности (для нашего примера таблица 2).

Таблица 2 – Таблица сопряженности

Фактические значения y_i	Предсказанные значения $\hat{y}_i$		Точность прогноза
	$\hat{y}_i = 0$	$\hat{y}_i = 1$	
$y_i = 0$	n_{11}=11	n_{12}=2	
$y_i = 1$	n_{21}=3	n_{22}=4	(11+4)/20=0,75 или 75%

Для построения такой таблицы в R можно воспользоваться функцией table(), предварительно добавив новый столбец к таблице logit с расчетными значениями переменной disease:

```
logit$disease_fitted<-ifelse (model1$fitted.values<0.5, 0, 1)
table(logit$disease, logit$disease_fitted)
```

Результат:

```
   0  1
0 11  2
1  3  4
```

Можно воспользоваться функцией CrossTable() из пакета gmodels

```
library(gmodels)
CrossTable(logit$disease, logit$disease_fitted, prop.c=FALSE, prop.r=FALSE, dnn = c("фактические значения", "расчетные значения"))
```

В результате получаем таблицу, где в каждой ячейке третье число – это доля частоты от объема выборки.

```
Total Observations in Table:  20


                   | расчетные значения
фактические значения |         0 |         1 | Row Total |
---------------------|-----------|-----------|-----------|
                   0 |        11 |         2 |        13 |
                     |     0.397 |     0.926 |           |
                     |     0.550 |     0.100 |           |
---------------------|-----------|-----------|-----------|
                   1 |         3 |         4 |         7 |
                     |     0.737 |     1.719 |           |
                     |     0.150 |     0.200 |           |
---------------------|-----------|-----------|-----------|
        Column Total |        14 |         6 |        20 |
---------------------|-----------|-----------|-----------|
```

Получается, что 0,550+0,200=0,750 – доля правильно классифицированных наблюдений, а 0,100+0,150=0,250 – доля неправильно классифицированных.

Если речь идет о сравнении нескольких альтернативных моделей бинарного выбора с разным количеством объясняющих переменных, то сравнивать качество альтернативных моделей можно, опираясь на значения информационных критериев Акаике (AIC) , Шварца (SC) и Хеннана–Куинна:

$$AIC = -2LnL_k + \frac{2p}{n},$$

$$SC = -2\frac{LnL_k}{n} + p\frac{\ln p}{n},$$

$$HQ = -2\frac{LnL_k}{n} + 2p\frac{\ln(Ln\,n)}{n},$$

здесь L_k – максимальное значение функции правдоподобия для k -й из альтернативных моделей, а p – количество объясняющих переменных в этой модели. При этом среди нескольких альтернативных моделей выбирается та, которая **минимизирует** значение статистики критерия. Заметим, что эти три критерия различаются размерами “штрафа”, который приходится платить за включение в модель большего количества объясняющих переменных.

Значение критерия AIC для нашей модели приведено в отчете функции summary().

ROC-анализ, показатель AUC

Построенная логит-модель возвращает вероятности того, что y=1. Но чтобы оценить точность модели, переходим от вероятностей к 0 и 1. Для этого выбирают порог — если значение вероятности выше него, мы будем считать, что модель предсказала 1, если ниже, то 0.

Значение порога зависит от многих факторов и будет влиять на качество модели. Прежде всего стоит ориентироваться на сферу деятельности, в которой вы проводите анализ. Если у вас качественные чистые данные и вам важна высокая точность, то и порог для предсказаний должен быть высокий — ≥0.9. Если же вы знаете, что вы работаете с зашумлёнными данными, и цена ошибки не так высока, то можете выбрать более либеральный критерий — 0.7−0.8. Самый либеральный критерий из возможных — 0.5, что по сути есть вероятность случайного угадывания. В нашем примере, когда добавляли к таблице logit столбец с расчетными значениями disease, в качестве порогового значения брали 0,5.

Для того, чтобы подобрать лучшее значение порога, для оценки точности модели, используют понятие ROC-кривой и ROC-анализа, а также AUC.

Как строится ROC-кривая:

1) Упорядочивают объекты по убыванию значения предсказанной вероятности и добавляют столбец истинных значений (0 и 1):

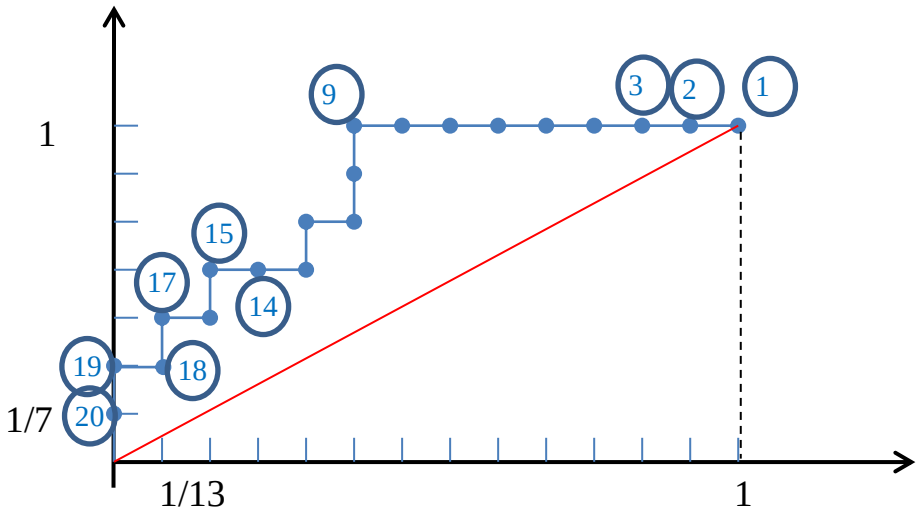
	age	disease	fitted_prob
20	84	1	0.77768244
19	81	1	0.74103277
18	80	0	0.72797787
17	79	1	0.71451840
16	72	0	0.61033933
15	68	1	0.54510568
14	62	0	0.44501779
13	60	0	0.41223493
12	59	1	0.39611243
11	50	0	0.26419173
10	49	1	0.25138287
9	44	1	0.19371788
8	42	0	0.17365484
6	41	0	0.16425561
7	41	0	0.16425561
5	32	0	0.09713217

	age	disease	fitted_prob
4	31	0	0.09141660
3	30	0	0.08600532
2	29	0	0.08088583
1	25	0	0.06307990

Showing 1 to 16 of 20 entries, 4 total columns

Если модель идеально справляется с предсказаниями, то в упорядоченном по значению предсказанной вероятности наборе данных сначала будут идти все наблюдения с истинным значением 1, а потом с 0.

2) Строят ROC-кривую, которая будет описывать качество нашей модели:
   a) Стартуют из точки (0,0) и конечная точка (1,1);
   b) Ось Oy делим на равные части, число которых равно количеству фактических 1 (у нас их 7, поэтому шаг равен 1/7≈0,14).
   c) Ось Ox делим на равные части, число которых равно количеству фактических 0 (у нас их 13, шаг равен 1/13=0,077).
   d) Идём по нашим данным сверху вниз, и когда встречаем наблюдение со значением 1, поднимаемся на графике на одно деление вверх; когда встречаем наблюдение со значением 0, сдвигаемся на одно деление вправо.
   e) В итоге получится такая кривая:



Каждая точка на этой кривой соответствует некоторому порогу вероятности отсечения объектов. Например, точка, соответствующая наблюдению 15 на рисунке, соответствует порогу вероятности 0.5 и имеет координаты (2/13, 4/7). Это означает, что если при переходе к предсказаниям модели мы будем использовать порог 0.5, то доля правильно классифицированных единиц (то есть когда фактически имеется заболевание и по модели оно тоже имеется) будет равна 4/7, а доля ситуаций, когда фактически наблюдались нули (то есть отсутствовало заболевание), а модель нам предсказала 1 (то есть, что заболевание имеется) окажется равной 2/13. Первая доля называется долей True Positive (TP), а вторая – долей False Positive (FP, ложноположительные предсказания, ошибочно предсказанные единицы).

В случае идеального упорядочивания наблюдений по предсказанной вероятности площадь под кривой (**ROC-AUC**, *area under a curve*) будет равна единице. Чем ближе значение ROC-AUC к единице, тем модель работает лучше. Значение 0.5 указывает на то, что модель совсем не ухватывает закономерность и точность её предсказаний на уровне случайного угадывания — ROC-кривая в этом случае будет проходить близко к диагонали, проведенной через точки с координатами (0, 0), (1, 1).
Для проведения ROC-анализа можно воспользоваться функциями roc() и auc() из пакета pROC. Первой функции на вход подается вектор истинных значений и вектор предсказанных вероятностей. Вторая же хочет получить результат работы первой.

```
install.packages("pROC")
library(pROC)
#Проведем ROC-анализ
ROC<-roc(logit$disease, model1$fitted.values)
#Рисуем ROC-кривую
plot(ROC)
#Определяем площадь под ROC-кривой: AUC
auc(ROC)
```

Результат выполнения:

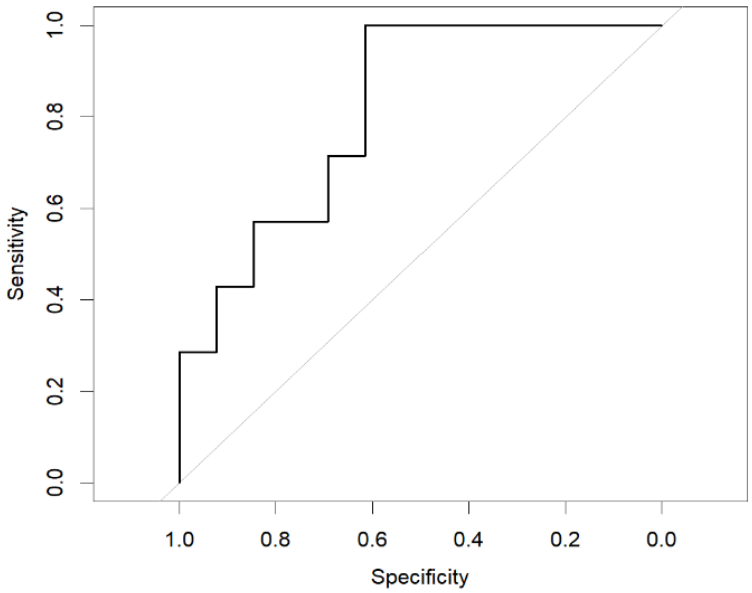


Рисунок. ROC-кривая

AUC=0.8132.