Множественная регрессия

Y = 7.2+3costs+4.5promotion+0.4books

R_adj = 0.78

3 – насколько изменится ожидаемое значение кассовых сборов фильма при единичном изменении costs при условии, что все остальные независимые переменные не изменяются

4.5 –

0.4 –

7.2 –

Значение исправленного коэффициента детерминации – 78% дисперсии зависимой переменной (y) объяснена полученной моделью

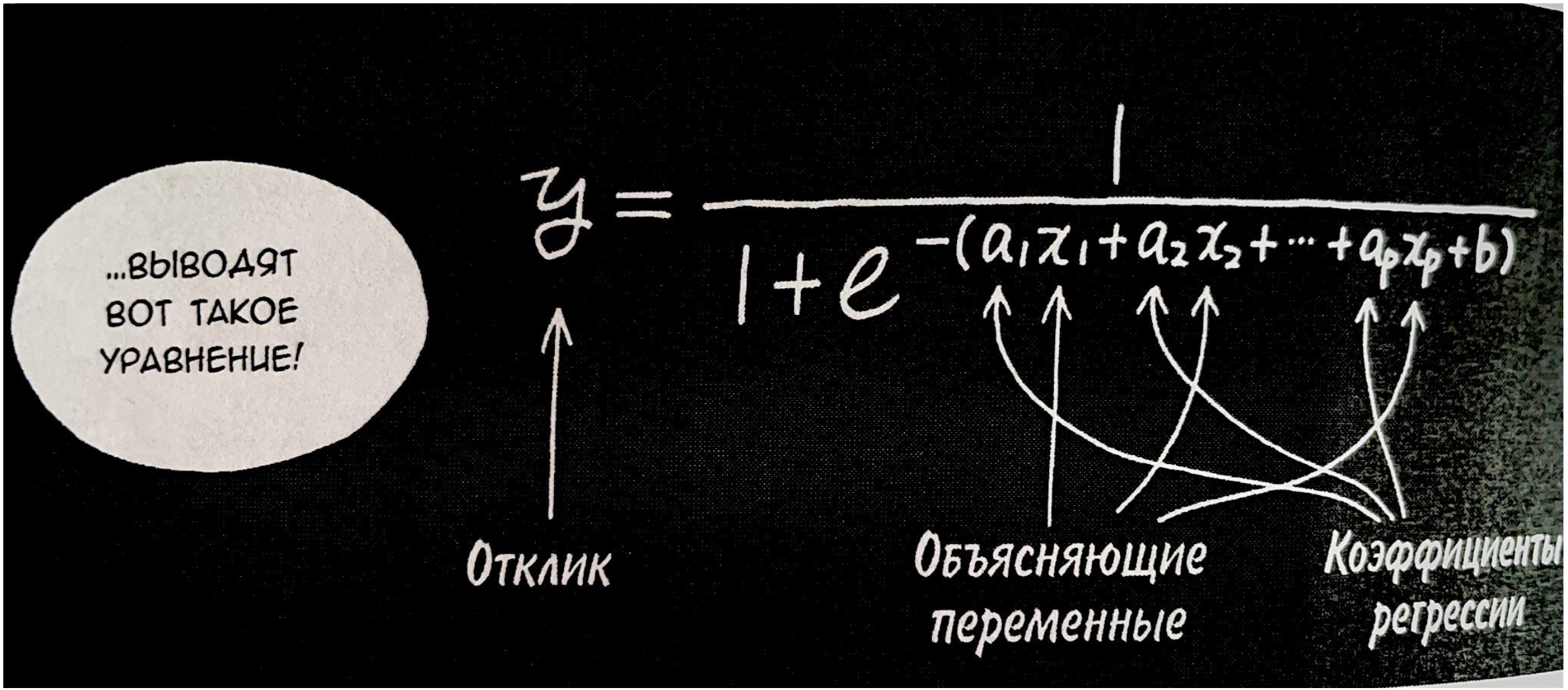
Логистическая регрессия

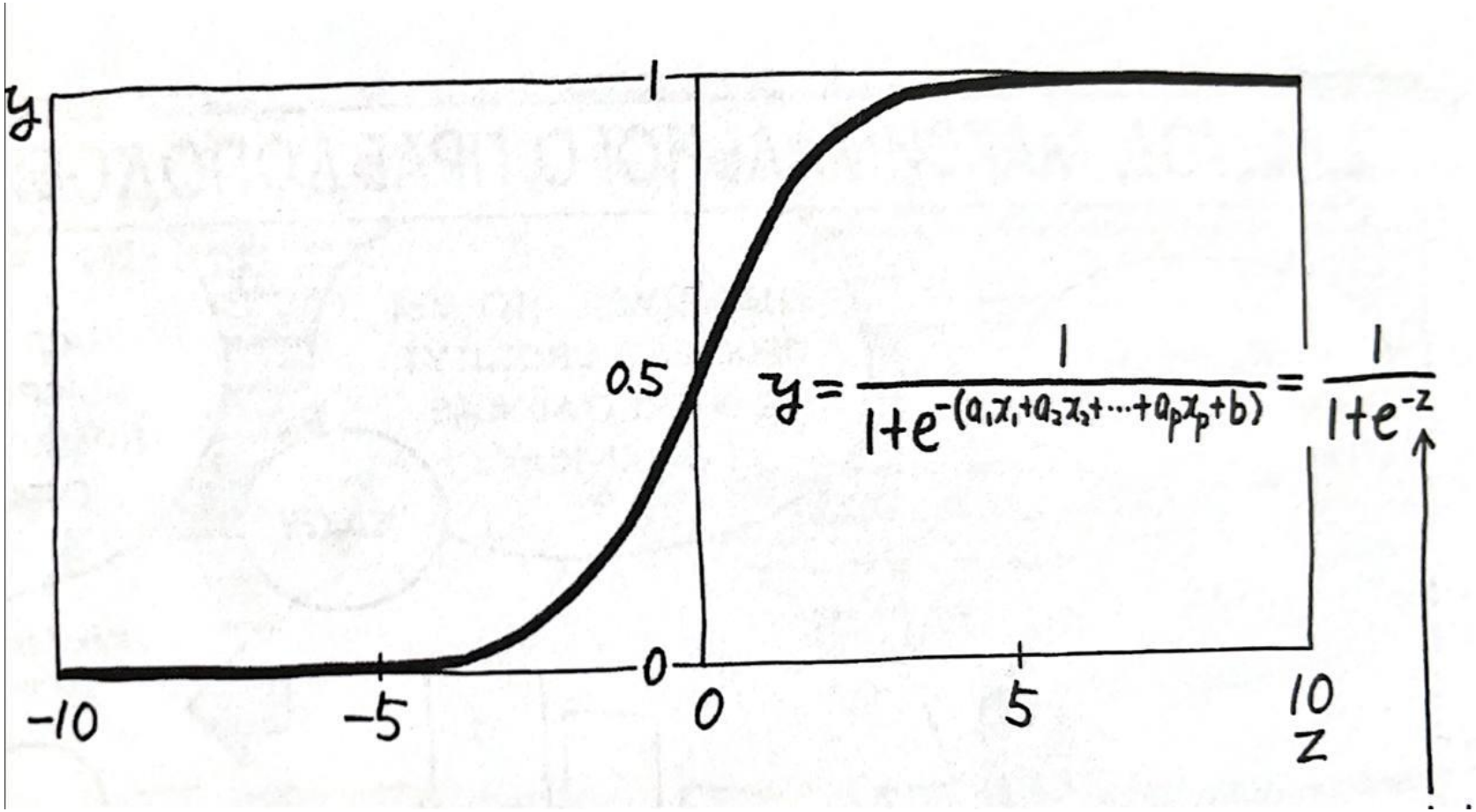
Изученные ранее регрессионный анализ и множественная регрессия были методами анализа для прогноза числовых значений: количество заказов чая со льдом, выручки магазина

Логистическая регрессия – прогноз вероятностей

Например, вероятность того, что абитуриент поступит в университет

Ни регрессионный анализ, ни множественная регрессия не дают возможности ограничить ожидаемые значения интервалом от 0 до 1


...ВЫВОДЯТ ВОТ ТАКОЕ УРАВНЕНИЕ!
$y = \frac{1}{1+e^{-(a_1x_1+a_2x_2+\cdots+a_px_p+b)}}$
Отклик
Объясняющие переменные
Коэффициенты регрессии

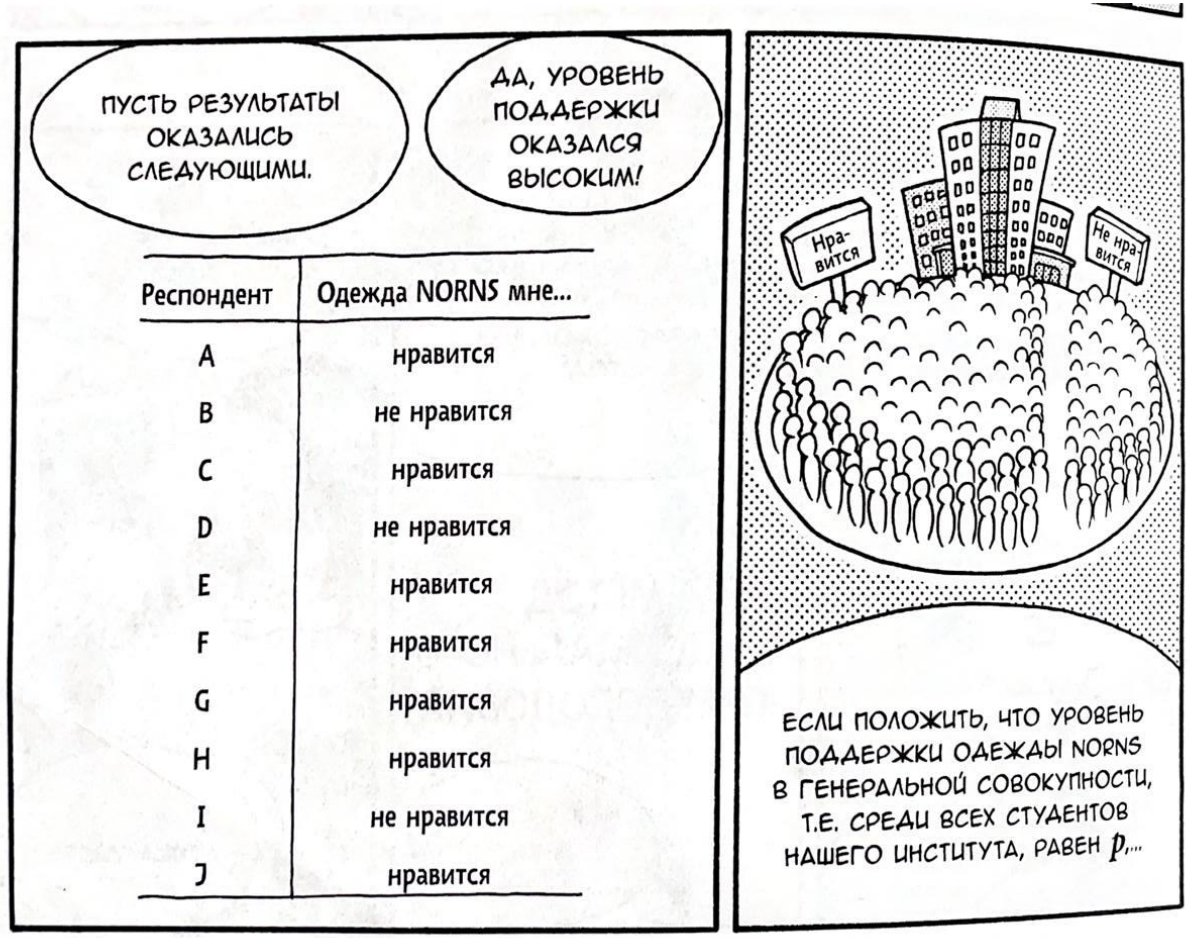


значения y находится между 0 и 1 при любых значениях z

$$z=a_1x_1+a_2x_2+\ldots+a_px_p+b!$$

Метод максимального правдоподобия

Пример. Спросили 10 студентов нравится или не нравится им форма



Вероятность получить приведенную в таблице картину окажется равной:
$p*(1-p)*p*(1-p)*p*p*p*p*(1-p)*p=p^7(1-p)^3$

Значение p равное уровню поддержки формы в генеральной совокупности “всех студентов нашего института”

$p^7(1-p)^3$ – функция правдоподобия
$\log\{p^7(1-p)^3\}$ – логарифмическая функция правдоподобия

Значения p, при котором **и функция правдоподобия, и логарифмическая функция** правдоподобия принимают максимальное значения - **называется оценка максимального правдоподобия**

Оценка максимального правдоподобия для примера

Шаг 1

Выводим функцию правдоподобия:

$$p \cdot (1-p) \cdot p \cdot (1-p) \cdot p \cdot p \cdot p \cdot p \cdot (1-p) \cdot p = p^7(1-p)^3$$

Шаг 2

Пишем логарифмическую функцию правдоподобия и упрощаем её:

$$L = \log\{p^7 (1-p)^3\}$$
$$= \log p^7 + \log(1-p)^3$$
$$= 7 \log p + 3 \log(1-p)$$

Шаг 3

Дифференцируем логарифмическую функцию правдоподобия L по p и приравниваем производную к 0:

$$\frac{dL}{dp} = 7 \cdot \frac{1}{p} + 3 \cdot \frac{1}{1-p} \cdot (-1) = 7 \cdot \frac{1}{p} - 3 \cdot \frac{1}{1-p} = 0$$

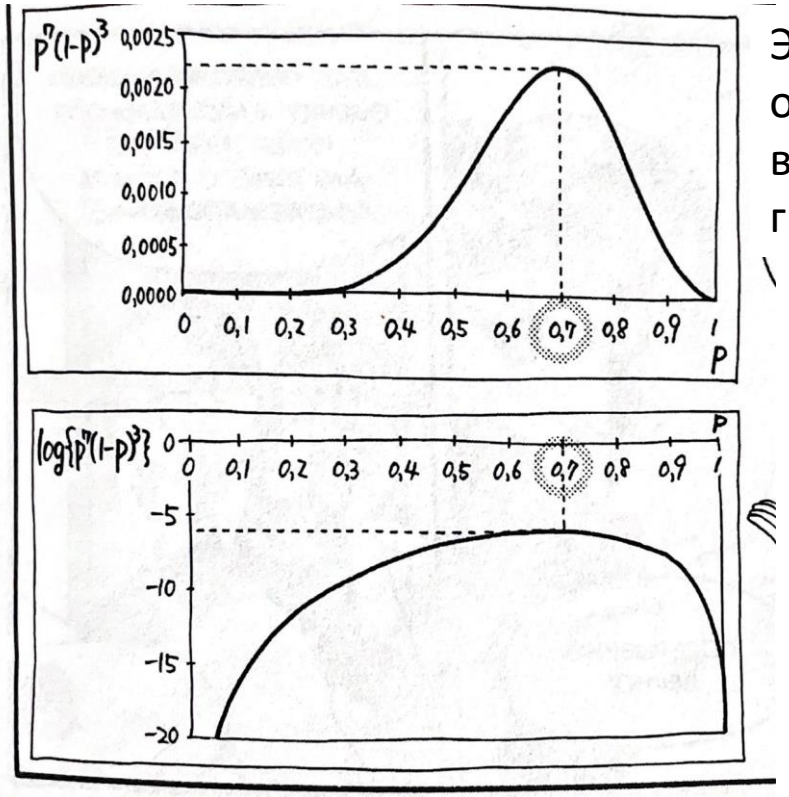
Шаг 4

Упрощаем выражение, полученное на шаге 3, и находим оценку максимального правдоподобия:

$$7 \cdot \frac{1}{p} - 3 \cdot \frac{1}{1-p} = 0$$

Домножаем обе части на $p(p-1)$:

$$\left(7 \cdot \frac{1}{p} - 3 \cdot \frac{1}{1-p}\right) \cdot p(1-p) = 0 \cdot p(1-p)$$
$$7(1-p) - 3p = 0$$
$$7 - 7p - 3p = 0$$
$$7 - 10p = 0$$
$$p = \frac{7}{10}$$

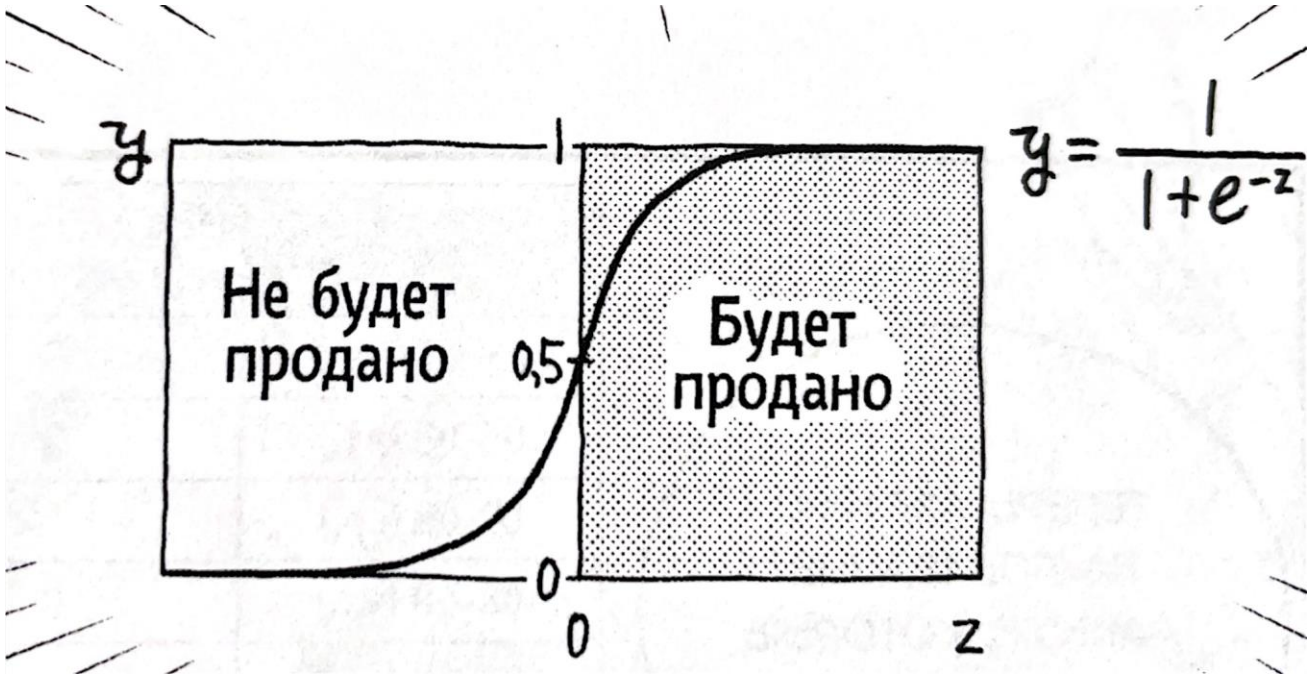


Это означает найти точку на оси x, которой соответствует вершина горки этих графиков

Пример в эксель

Трактовка отклика

Покупка мороженого по спец меню (1-продано, 0 – не продано)



Логистическая регрессия

- Только измеряемые данные
- Только неизмеряемые данные
- Комбинация измеряемых и неизмеряемых данных

Процесс построения логистической регрессии

1.Проверка целесообразности проведения логистической регрессии с помощью точечных графиков всех объясняющих переменных и отклика

2.Вывод уравнения логистической регрессии

3.Проверка точности уравнения логистической регрессии

4.Проведение ‘проверки значимости коэффициентов регрессии’

5.Построение прогноза

1. Процесс построения логистической регрессии

Рассчитываем коэффициент корреляции между x1 и y и x2 и y

Коэффициент корреляции между x1 и y	0,509524665
Коэффициент корреляции между x2 и y	0,48280455

где

x1	среда, суббота или воскресение
x2	максимальная температура
y	продалось ли мороженое по спец меню или нет

2.Вывод уравнения множественной регрессии

Шаг 1

Выполняем вычисления согласно приведённой ниже таблице.

	Среды, субботы или воскресенья x_1	Максимальная температура x_2	Картина продаж спецменю NORNS y	Картина продаж спецменю NORNS $\hat{y} = \frac{1}{1 + e^{-(a_1 x_1 + a_2 x_2 + b)}}$
05-08 (пон)	0	28	1	$\frac{1}{1 + e^{-(a_1 \cdot 0 + a_2 \cdot 28 + b)}}$
06-08 (втр)	0	24	0	$\frac{1}{1 + e^{-(a_1 \cdot 0 + a_2 \cdot 24 + b)}}$
:	:	:	:	:
25-08 (вск)	1	24	1	$\frac{1}{1 + e^{-(a_1 \cdot 1 + a_2 \cdot 24 + b)}}$

Шаг 2

Записываем функцию правдоподобия:

$$\underbrace{\frac{1}{1+e^{-(a_1\cdot 0+a_2\cdot 28+b)}}}_{\text{продано}}\cdot\underbrace{\left(1-\frac{1}{1+e^{-(a_1\cdot 0+a_2\cdot 24+b)}}\right)}_{\text{не продано}}\cdot\ldots\cdot\underbrace{\frac{1}{1+e^{-(a_1\cdot 1+a_2\cdot 24+b)}}}_{\text{продано}}$$

Шаг 3

Записываем логарифмическую функцию правдоподобия L:

$$L=\log\left\{\frac{1}{1+e^{-(a_1\cdot 0+a_2\cdot 28+b)}}\cdot\left(1-\frac{1}{1+e^{-(a_1\cdot 0+a_2\cdot 24+b)}}\right)\cdot\ldots\cdot\frac{1}{1+e^{-(a_1\cdot 1+a_2\cdot 24+b)}}\right\}$$

$$=\log\left(\frac{1}{1+e^{-(a_1\cdot 0+a_2\cdot 28+b)}}\right)+\log\left(1-\frac{1}{1+e^{-(a_1\cdot 0+a_2\cdot 24+b)}}\right)+\ldots+$$

$$+\log\left(\frac{1}{1+e^{-(a_1\cdot 1+a_2\cdot 24+b)}}\right)$$

Шаг 4

Находим оценку максимального правдоподобия.

Оценка максимального правдоподобия, т.е. значения a_1, a_2, b, при которых логарифмическая функция правдоподобия L имеет максимальное значение:

$$\begin{cases} a_1 = 2{,}44 \\ a_2 = 0{,}54 \\ b = -15{,}20 \end{cases}$$

Шаг 5

Записываем уравнение логистической регрессии, которое, согласно шагу 4, будет иметь вид:

$$y = \frac{1}{1 + e^{-(2{,}44x_1 + 0{,}54x_2 - 15{,}20)}}$$

3.Проверка точности уравнения логистической регрессии

$$R^2 = 1 - \frac{\text{Макс. значение логарифмич. функции правдоподобия } L}{n_1 \log n_1 + n_0 \log n_0 - (n_1 + n_0) \log (n_1 + n_0)}$$

n_1	Число экземпляров, для которых значение отклика = 1
n_0	Число экземпляров, для которых значение отклика = 0

$$R^2 = 1 - \frac{\text{Макс. значение логарифмич. функции правдоподобия } L}{n_1 \log n_1 + n_0 \log n_0 - (n_1 + n_0) \log (n_1 + n_0)}$$

$$= 1 - \frac{-8,9}{8 \log 8 + 13 \log 13 - (8 + 13) \log (8 + 13)}$$

$R^2 = 0.3622$

Чем больше точность уравнения логистической регрессии, тем ближе он к 1, в противном случае – к 0.

Notes: считается, что коэффициент детерминации уравнения логистической регрессии не склонен принимать большие значения, поэтому его просто принять к сведению

Относительная ошибка дискриминации

Число экземпляров с несовпадением фактического и ожидаемого значений/Общее число экземпляров

Чем меньше указанное значение, тем **точнее** уравнение логистической регрессии

Проверка значимости

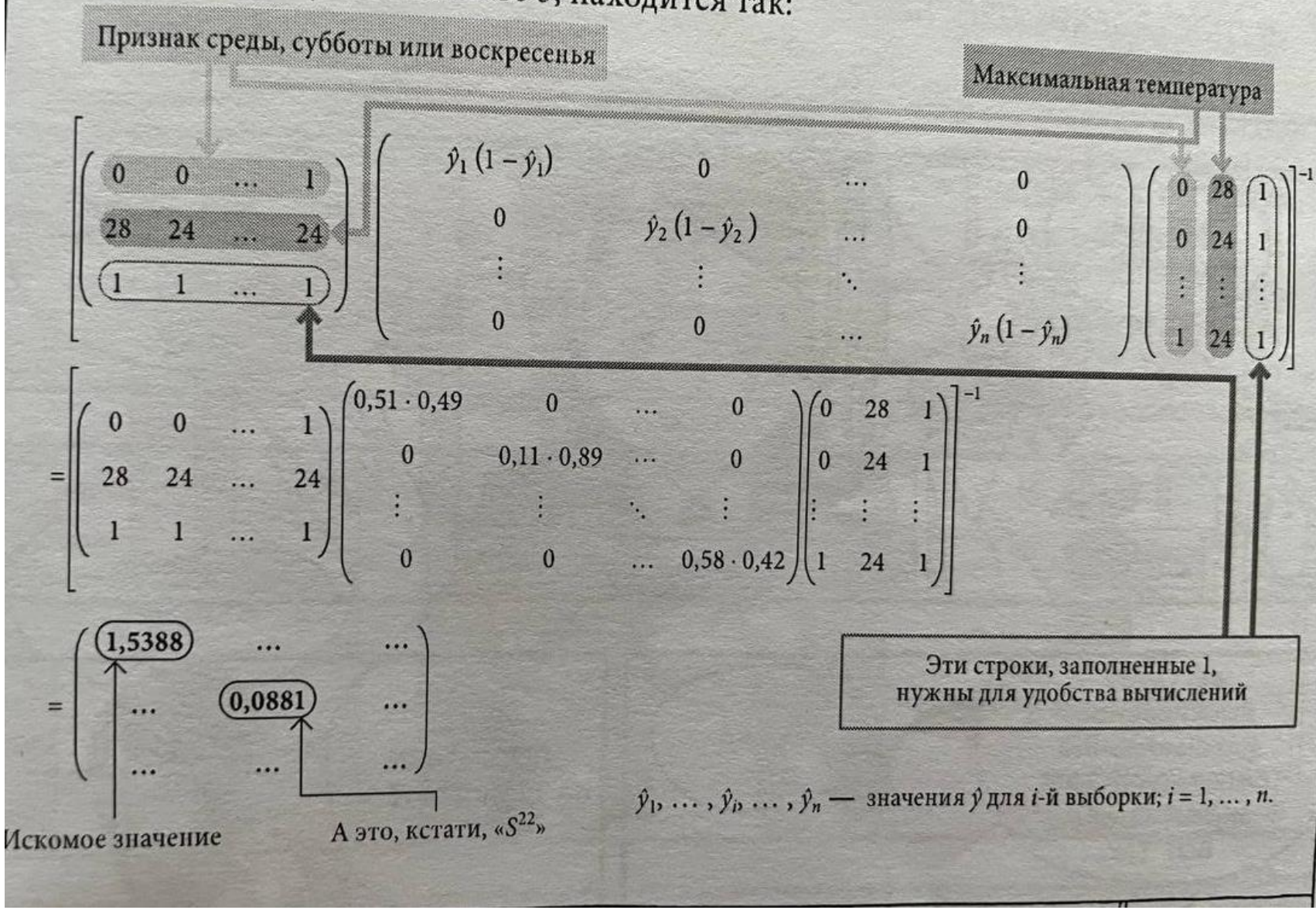
Начинаем с «совместной проверки значимости коэффициентов регрессии»! Кстати, оценку путём нижеприведённых вычислений обычно называют **тестом отношения правдоподобия.**

Шаг 1	Определение генеральной совокупности.	Определяем генеральную совокупность как «дни с признаком среды, субботы или воскресенья x_1 и с максимальной температурой x_2 °C».
Шаг 2	Построение нулевой гипотезы и альтернативной гипотезы.	Нулевая гипотеза: $A_1=A_2=0$ Альтернативная гипотеза: $A_1=A_2=0$ не выполняется
Шаг 3	Выбор вида статистической проверки.	Будем проводить «совместную проверку значимости коэффициентов регрессии».
Шаг 4	Назначение уровня значимости.	Выбираем уровень значимости равным 0,05.
Шаг 5	Нахождение значения статистического критерия по данным выборки.	Мы собираемся провести «совместную проверку значимости коэффициентов регрессии», в котором значение статистического критерия вычисляется по формуле: $2\cdot(L_{max}-n_1\ln n_1-n_0\ln n_0+(n_1+n_0)\ln(n_1+n_0))$ где L_{max} - максимальное значение логарифмической функции правдоподобия В нашем примере это значение равно: $2\cdot(-8{,}9010-8\ln 8-13\ln 13+(8+13)\ln(8+13))=10{,}1$. Кроме того, в нашем примере в случае верности нулевой гипотезы статистический критерий будет подчиняться распределению хи-квадрат с числом степеней свободы, равным 2 (т.е. числу объясняющих переменных).
Шаг 6	Сравнение значения P, которое соответствует значению статистического критерия, найденному в шаге 5, с уровнем значимости.	Уровень значимости равен 0,05. Значение P, которое соответствует значению статистического критерия 10,1, равно 0,006. $0{,}0006<0{,}05$, т.е. значение P ниже уровня значимости.
Шаг 7	Если сравнение на шаге 6 показало, что значение P ниже уровня значимости, то делается вывод «альтернативная гипотеза правильна». В противном случае делается вывод «нулевая гипотеза не может быть признана ошибочной».	Значение P оказалось ниже уровня значимости. Следовательно верна альтернативная гипотеза, согласно которой $A_1=A_2=0$ не выполняется.

А теперь — раздельную проверку значимости коэффициентов! Мы попробуем силы на a_1! Кстати, проверка по приведённой ниже методике называется *тестом Вальда.*

Шаг 1	Определение генеральной совокупности.	Определяем генеральную совокупность как «дни с признаком среды, субботы или воскресенья x_1, и с максимальной температурой x_2 °C».
Шаг 2	Построение нулевой гипотезы и альтернативной гипотезы.	Нулевая гипотеза: $A_1=0$ Альтернативная гипотеза: $A_1\neq 0$
Шаг 3	Выбор вида статистической проверки.	Будем проводить раздельную проверку значимости коэффициентов регрессии.
Шаг 4	Назначение уровня значимости.	Назначаем уровень значимости равным 0,05.
Шаг 5	Нахождение значения статистического критерия по данным выборки.	Мы собираемся провести «раздельную проверку значимости коэффициентов регрессии», в котором значение статистического критерия вычисляется по формуле: $\frac{a_1^2}{S^{11}}$. В нашем примере это значение равно: $\frac{2{,}44^2}{1{,}5388}=3{,}9$ Кроме того, в нашем примере в случае верности нулевой гипотезы статистический критерий будет подчиняться распределению хи-квадрат числом степеней свободы, равным 1.
Шаг 6	Сравнение значения P, которое соответствует значению статистического критерия, найденному в шаге 5, с уровнем значимости.	Уровень значимости равен 0,05. Значение P, которое соответствует значению статистического критерия 3,9, равно 0,0489. $0{,}0489<0{,}05$, т.е. значение P ниже уровня значимости.
Шаг 7	Если сравнение на шаге 6 показало, что значение P ниже уровня значимости, то делается вывод: «Альтернативная гипотеза правильна», иначе: «Нулевая гипотеза не может быть признана ошибочной».	Значение P оказалось ниже уровня значимости. Следовательно верна альтернативная гипотеза, согласно которой $A_1\neq 0$.

«S^{11}», используемое на шаге 5, находится так:



$\hat{y}_1, \dots, \hat{y}_i, \dots, \hat{y}_n$ — значения $\hat{y}$ для i-й выборки; $i = 1, \dots, n$.

Построение прогноза

подставить новые значения x1 и x2 и рассчитать значение *y*

если полученное значение *y* получилось меньше, чем 0.5 – значит y=0, то есть мороженое из спец меню не продастся, если больше 0.5 – то y=1, мороженое по спец меню продастся