

(**table_booking** станет 1, вместо 0) размер чека в нем тоже может вырасти (**avg_receipt**), или что если цены в ресторане вырастут (**avg_receipt** станет больше), то может появиться возможность бронировать столик (**table_booking** станет 1, вместо 0).

✓ Глава 8. Линейная регрессия

✓ 8.1 Что такое регрессия?

Регрессия представляет собой мощный статистический инструмент, позволяющий выявлять и анализировать взаимосвязи между явлениями. Эта методика незаменима, когда необходимо установить, каким образом изменение одной величины влияет на другую величину. Регрессия помогает исследователю строить прогнозы, оценивать последствия различных воздействий и находить скрытые закономерности в данных. Применение регрессионного анализа широко востребовано в маркетинге, социологии и многих других областях науки и практики, обеспечивая глубокое понимание сложных процессов и принятие обоснованных решений.


Регрессия - уравнение, которое показывает как один или несколько факторов оказывают влияние на другой **целевой** признак.

Предположим, у нас есть выборка по семи ресторанам. Про каждый из этих ресторанов нам известен средний размер чека (**avg_receipt**), рейтинг (**rate**) и местоположение (**area**; 0 - центр, 1 - не центр):

```
import pandas as pd
```

```
df = pd.DataFrame()  
df['avg_receipt'] = [50, 100, 200, 200, 150, 350, 350]
```

```
df['rate'] = [0.5, 1.5, 1.5, 2.5, 3.5, 3.5, 4.5]
df['area'] = [1, 1, 0, 0, 0, 0, 0]
df
```



	avg_receipt	rate	area
0	50	0.5	1
1	100	1.5	1
2	200	1.5	0
3	200	2.5	0
4	150	3.5	0
5	350	3.5	0
6	350	4.5	0

РИС 27

Мы хотим выяснить, как рейтинг ресторана (**rate**) влияет на средний размер чека (**avg_receipt**) в нем. Запишем это в чуть более математической форме:

$$\widehat{\text{avg_receipt}} = w_0 + w_1 * \text{rate}$$

Мы хотим предсказать размер чека (**avg_receipt**). С точки зрения регрессии этот признак называется **целевой** или **зависимой переменной**. В уравнении линейной регрессии целевой переменной может выступать только **количественный признак**.

Рейтинг ресторана (**rate**) - фактор, оказывающий влияние на размер чека (**avg_receipt**). В модели регрессии рейтинг (**rate**) - это **предиктор** или **независимая переменная**. В отличие от целевой, предиктором может быть выступать и количественный, и категориальный признак.

На вопрос, как рейтинг (**rate**) влияет на размер чека (**avg_receipt**), отвечают параметры (коэффициенты) регрессии w_0 и w_1 . w_0 называют **свободным коэффициентом**. Он

стоит в одиночестве, без пары, в отличие от w_1
 - **коэффициента независимой переменной**.

Регрессия, в которой только один предиктор (в нашем случае это рейтинг (**rate**)), называется **однофакторной**.

Но мы понимаем, что в реальной жизни не только рейтинг (**rate**) влияет на размер чека (**avg_receipt**). На него может влиять тип кухни, количество сотрудников и многое другое.

Такая регрессия, с более чем одним предиктором, называется **многофакторной**. В наших данных, помимо рейтинга, есть еще местоположение (**area**). Если мы хотим добавить этот фактор в модель, то с точки зрения математики регрессию можно записать так:

$$\widehat{\text{avg_receipt}} = w_0 + w_1 * \text{rate} + w_2 * \text{area}$$

rate, area - это независимые переменные регрессии.

w_1, w_2 - коэффициенты независимых переменных

w_0 - свободный коэффициент

	ОДНОФАКТОРНАЯ	МНОГОФАКТОРНАЯ
Уравнение регрессии	$\hat{Y} = w_0 + w_1 * X$	$\hat{Y} = w_0 + w_1 * X_1 + \dots + w_k * X_k$
Целевая (зависимая) переменная <i>только количественная</i>	\hat{Y}	\hat{Y}
Предикторы/независимые переменные	X	X_1, \dots, X_k
Параметры:		
Коэффициенты независимой переменной	w_1	w_1, \dots, w_k
Свободный коэффициент	w_0	w_0

✓ 8.2 Как обучить линейную регрессию?

После формализации модели нужно обучить ее. **Обучить модель** - значит рассчитать её параметры (w_0, \dots, w_k) . Параметры модели вычисляются по формулам, выведенным из **метода наименьших квадратов (МНК)**, или на английском ordinary least squares (OLS).

Формулы метода наименьших квадратов

Коэффициент независимой переменной (w_1):

$$w_1 = \frac{(x_1 - \bar{x}) * (y_1 - \bar{y}) + \dots + (x_n - \bar{x}) * (y_n - \bar{y})}{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}$$

x_1, \dots, x_n - каждое из значений предиктора

\bar{x} - среднее арифметическое предиктора

y_1, \dots, y_n - каждое значение целевой переменной

\bar{y} - среднее арифметическое целевой переменной

n - количество наблюдений (в наших данных 7 наблюдений)

Свободный коэффициент (w_0):

$$w_0 = \bar{y} - w_1 * \bar{x}$$

\bar{y} - среднее арифметическое целевой переменной

w_1 - коэффициент независимой переменной

\bar{x} - среднее арифметическое предиктора

Мы сконцентрируемся на том, как обучить модель с помощью библиотеки `statsmodels.api`. Первое, что нам нужно сделать - импортировать библиотеку; при импорте дадим ей короткий псевдоним `sm`:

```
import statsmodels.api as sm
```

Ошибка `ModuleNotFoundError: No module named 'statsmodels'`

Если при запуске кода вышла такая ошибка, это значит, что модуль Statsmodels еще не установлен на компьютере. Чтобы его установить, нужно перед импортом добавить строку:

```
!pip install statsmodels
```

В эту библиотеку уже зашиты формулы метода наименьших квадратов, поэтому нам не придется их учить, а нужно будет лишь написать несколько несложных команд, чтобы найти значения параметров:

1. Обозначаем целевую переменную:

```
Y = df['целевая']
```

2. Обозначем предиктор(ы) и добавляем константный признак с помощью функции `sm.add_constant()` для корректного расчета коэффициентов регрессии:

```
X = sm.add_constant(df['предиктор']) # однофа
```

или

```
X = sm.add_constant(df[['предиктор 1', 'предик
```

3. Обучаем (`.fit()`) линейную регрессию с помощью метода наименьших квадратов (`sm.OLS()`):

```
модель = sm.OLS(Y, X).fit()
```

4. Выводим параметры регрессии с помощью атрибута `.params`:

```
модель.params
```

0

const	w_0	свободный коэффициент
предиктор 1	w_1	коэффициент независимой переменной
предиктор 2	w_2	коэффициент независимой переменной
...

Для начала обучим однофакторную модель, которая только на основе рейтинга ресторана (**rate**) будет предсказывать средний размер чека (**avg_receipt**):

$$\widehat{\text{avg_receipt}} = w_0 + w_1 * \text{rate}$$

```
Y = df['avg_receipt'] # целевая переменная
X1 = sm.add_constant(df['rate']) # предиктор
```

```
model1 = sm.OLS(Y, X1).fit() #обучаем модель
model1.params # выводим параметры
```



0

const	33.333333
rate	66.666667

dtype: float64

$$\widehat{\text{avg_receipt}} = 33.33 + 66.67 * \text{rate}$$

Для обучения многофакторной регрессии, которая предскажет размер чека (**avg_receipt**) не только на основе рейтинга ресторана (**rate**), но и по местоположению (**area**), нужно написать следующий код:

$$\widehat{\text{avg_receipt}} = w_0 + w_1 * \text{rate} + w_2 * \text{area}$$

```
Y = df['avg_receipt'] # целевая переменная
X2 = sm.add_constant(df[['rate', 'area']]) # п
```

```
model2 = sm.OLS(Y, X2).fit() #обучаем модель
model2.params # выводим параметры
```



0

const 100.438596**rate** 48.245614**area** -73.684211**dtype:** float64

$$\widehat{\text{avg_receipt}} = 100.44 + 48.25 * \text{rate} - 73.68 * \text{area}$$

✓ 8.3 Как интерпретировать линейную регрессию?

В прошлом разделе мы обучили линейную регрессию, а теперь нужно выяснить, что значат эти параметры.

ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ

Рассмотрим уравнение однофакторной регрессии:

$$\widehat{Y} = w_0 + w_1 * X$$

$$\widehat{\text{avg_receipt}} = 33.33 + 66.67 * \text{rate}$$

Такая регрессия на самом деле является уравнением прямой, которое вы наверняка проходили в школьном курсе геометрии. Если мы нанесем ее на диаграмму рассеяния, то она будет описывать тренд наших данных:

РИС 28

Свободный коэффициент w_0 (33.33) - это место, где прямая пересекает ось Y , т.е. ось целевой переменной (avg_receipt)

Коэффициент независимой переменной w_1 (66.67) регулирует угол наклона прямой. Если он положительный, как в нашем случае, это значит, признаки связаны прямой

зависимостью, а если угол отрицательный, то обратной.

ПРАКТИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ

Теперь рассмотрим, как можно объяснить практически полученные нами коэффициенты, и как можно интерпретировать регрессию в целом.

Для начала рассмотрим однофакторную регрессию:

$$\widehat{\text{avg_receipt}} = 33.33 + 66.67 * \text{rate}$$

- **Интерпретация регрессии**

Предположим, что в городе появился новый ресторан. Мы знаем, что его рейтинг *3.5 балла (rate)*. И нам нужно выяснить, какой там средний размер чека. Все что нам нужно сделать, это вместо **rate** подставить 3.5 и посчитать:

$$33.33 + 66.67 * 3.5$$


$$266.675$$

Получается, что в ресторане с рейтингом 3.5 (**rate**) средний размер чека (**avg_receipt**) будет составлять 266.675 у.е..

РИС 29

- **Свободный коэффициент (w_0)**

В однофакторной регрессии свободный коэффициент (w_0) показывает, чему будет