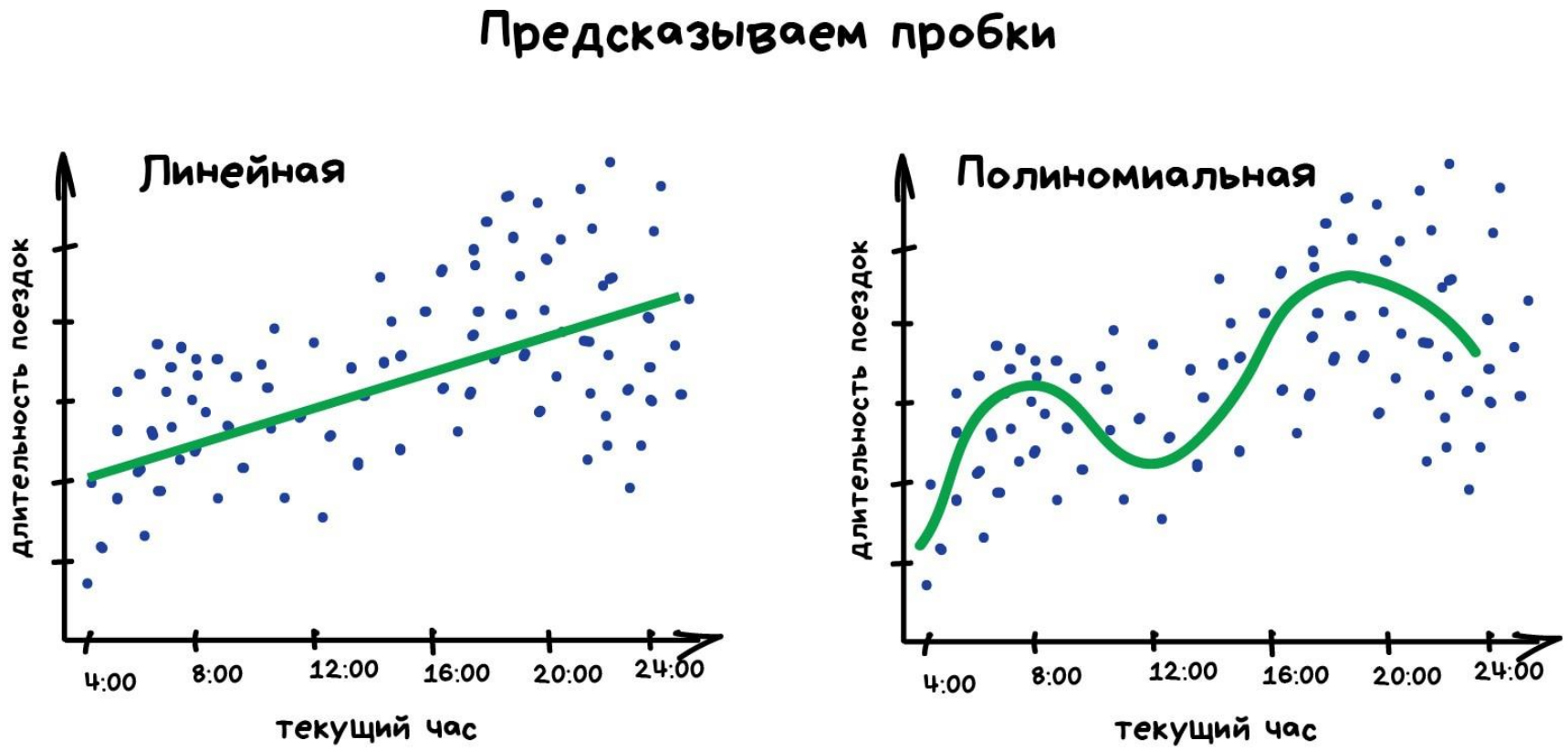
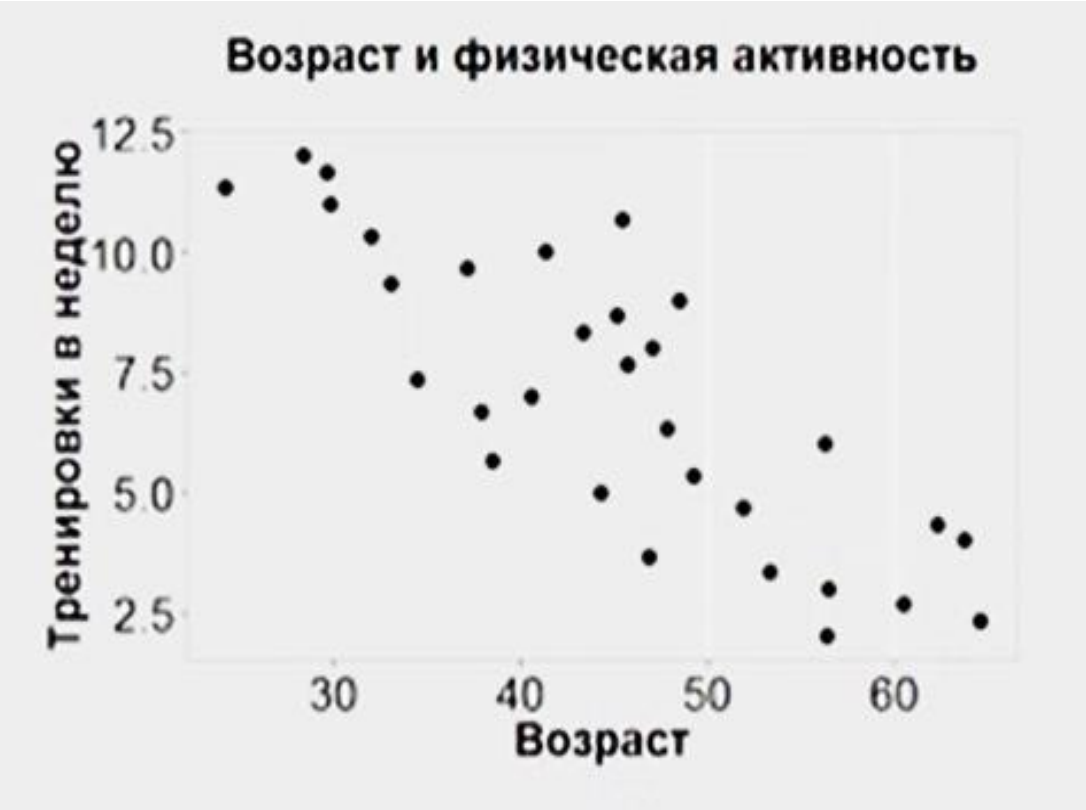
Линейная регрессия

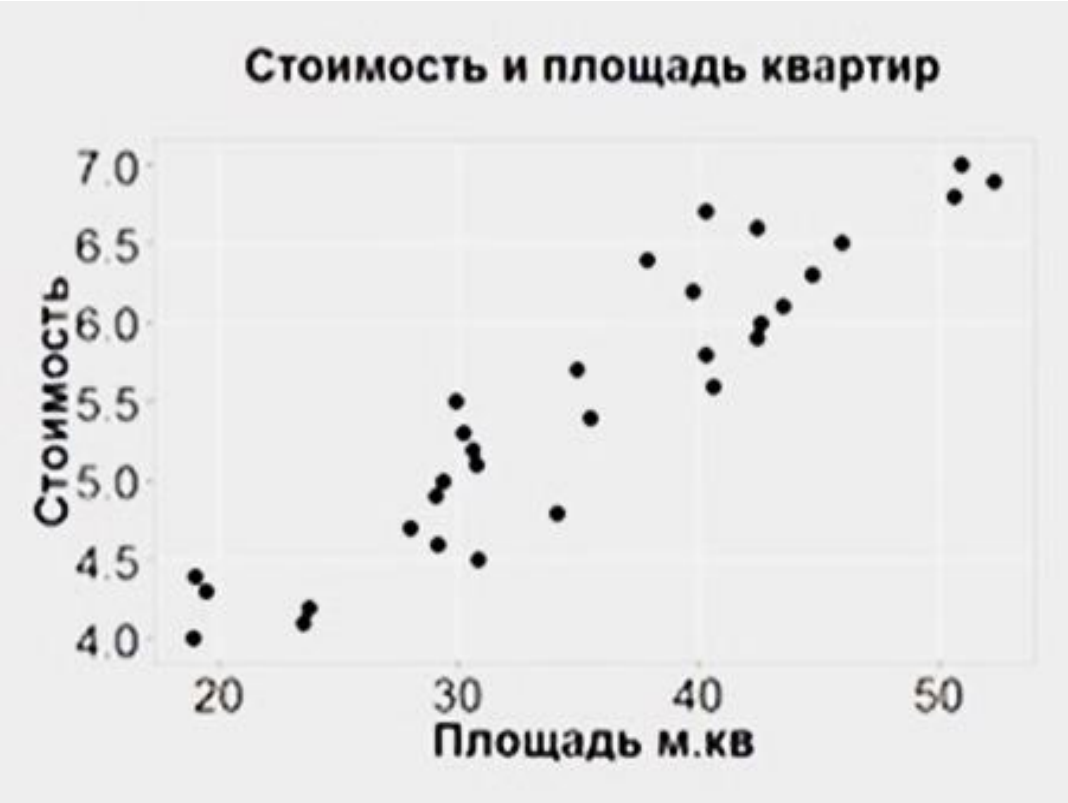
Линейная регрессия (Linear regression) — модель зависимости переменной *x* от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости.



Корреля́ция (от лат. correlatio «соотношение»)— статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Простыми словами **корреляция** – это взаимосвязь двух или нескольких случайных параметров. Когда изменения одной величины вызывают изменения другой.





Значительная корреляция между двумя случайными величинами всегда является свидетельством существования некоторой статистической связи в данной выборке, но эта связь не обязательно должна наблюдаться для другой выборки и иметь причинно-следственный характер.

Можно сказать, что **корреляция — это взаимосвязь без гарантий**

Например, чем выше уровень благосостояния человека, тем больше его продолжительность жизни. Однако нельзя с уверенностью сказать, что определенный миллионер проживет дольше конкретного нищего. **Это лишь статистическая вероятность, которая может не сработать для одного конкретного случая.** Этим корреляция отличается от линейной зависимости, где исход известен со 100-процентной вероятностью. Но если мы возьмем выборку из сотни тысяч богачей и такого же числа малоимущих, сравним их продолжительность жизни, то общая тенденция будет верна.

- **Корреляция двух величин может свидетельствовать о существовании общей причины, хотя сами явления напрямую не взаимодействуют.** Например, обледенение становится причиной как роста травматизма из-за падений, так и увеличения аварийности среди автотранспорта. В этом случае две величины (травматизм из-за падений пешеходов и аварийность автотранспорта) будут коррелировать, хотя они не связаны причинно-следственно друг с другом, а лишь имеют стороннюю общую причину — гололедицу.
- В то же время, **отсутствие корреляции между двумя величинами ещё не значит, что между ними нет никакой связи**. Например, зависимость может иметь сложный нелинейный характер, который корреляция не выявляет.

Коэффициент корреляции

Математической мерой корреляции двух случайных величин служит **коэффициент корреляции**.

Он отражает силу и направление взаимосвязи величин и находится в промежутке от -1 до 1.

$$r_{xy} = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

где $\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$, $\overline{Y} = \frac{1}{n}\sum_{t=1}^{n} Y_t$ — среднее значение выборок.

Словесная интерпретация величин коэффициента корреляции:

Значение коэффициента корреляции r	Интерпретация
0 < r <= 0,2	Очень слабая корреляция
0,2 < r <= 0,5	Слабая корреляция
0,5 < r <= 0,7	Средняя корреляция
0,7 < r <= 0,9	Сильная корреляция
0,9 < r <= 1	Очень сильная корреляция

Ш У Т К А

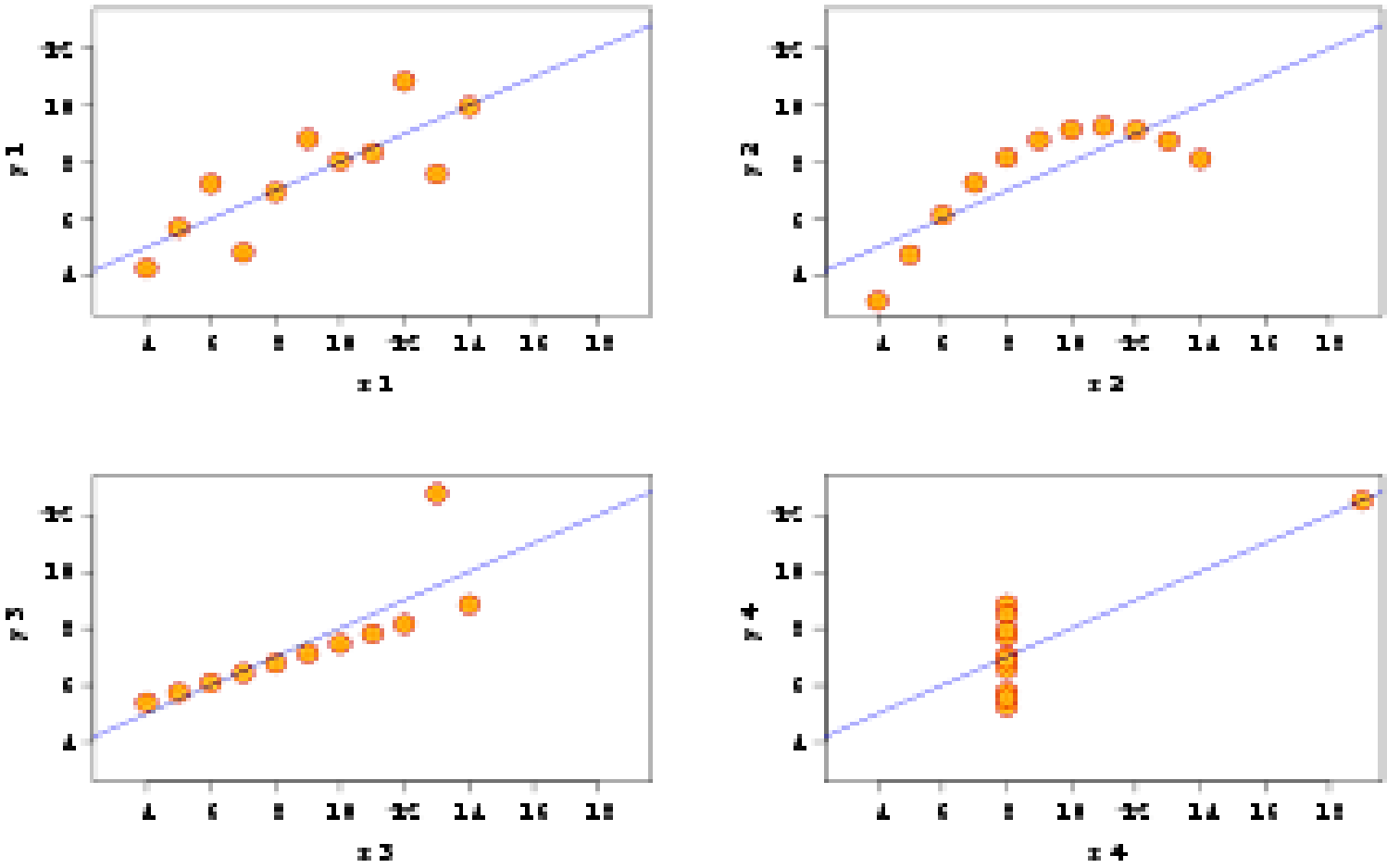
Значение	Какая корреляция?	О чем это говорит?
r = 1	Сильная положительная корреляция	Люди, которые едят чернику, обладают острым зрением. Ешьте чернику!
r < 0,5	Слабая положительная корреляция	Некоторые люди, которые любят чернику, обладают острым зрением. Но это не точно. Короче, ничего не пока понятно. Но лучше есть чернику на всякий случай.
r = 0	Корреляция отсутствует	Черника и зрение никак не связаны.
r < - 0,5	Слабая отрицательная корреляция	Бывают случаи ухудшения зрения из-за черники. Не стоит рисковать.
r = - 1	Сильная отрицательная корреляция	Практически все, кто ел чернику, ослепли. Берегитесь черники!

Для непрерывных переменных применяется коэффициент корреляции Пирсона.

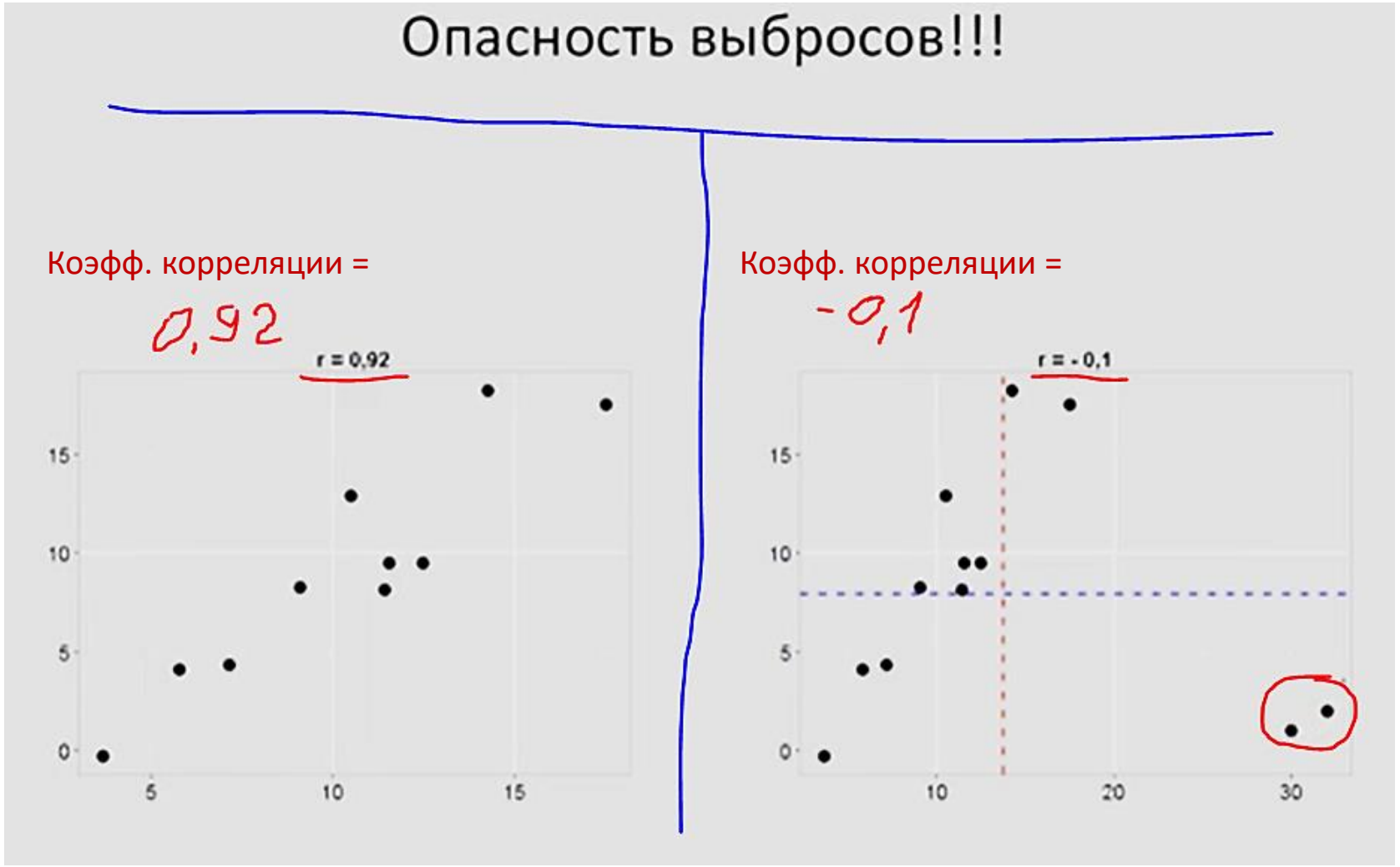
Ограничения коэффициентов корреляции Пирсона

1. **Характер связи должен быть линейным и монотонным**, в противном случае применение коэффициента корреляции Пирсона будет некорректным:

Пример: четыре различных набора данных, **коэффициент корреляции на которых равен 0.81**



2. **Должно обеспечиваться предположение о нормальности распределения**, любые выбросы, ассиметрии и прочие нарушения предположения о нормальности плохо влияют на величину коэффициента корреляции:



При наличии выбросов лучше использовать **коэффициент корреляции Спирмена**

Коэффициент ранговой корреляции Спирмена [править | править код]

Степень зависимости двух случайных величин (признаков) X и Y может характеризоваться на основе анализа получаемых результатов $(X_1, Y_1), \ldots, (X_n, Y_n)$. Каждому показателю X и Y присваивается ранг. Ранги значений X расположены в естественном порядке $i = 1, 2, \ldots, n$. Ранг Y записывается как R_i и соответствует рангу той пары (X, Y), для которой ранг X равен i. На основе полученных рангов X_i и Y_i рассчитываются их разности d_i и вычисляется коэффициент корреляции Спирмена:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

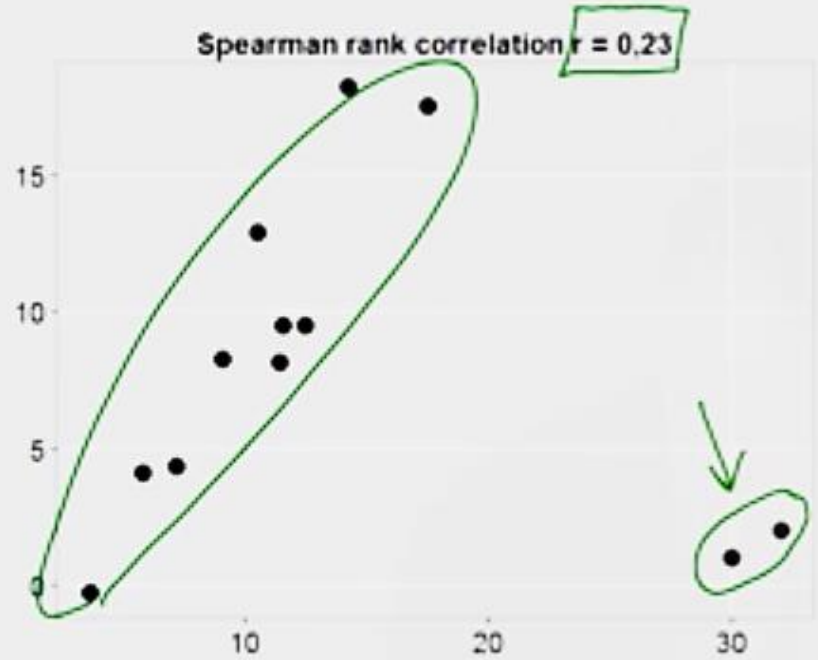
Значение коэффициента меняется от −1 (последовательности рангов полностью противоположны) до +1 (последовательности рангов полностью совпадают). Нулевое значение показывает, что признаки независимы.

Итог для выборки с выбросами

коэффициент корреляции Спирмена = 0.23

коэффициент корреляции Пирсона = -0.1

Коэффициент корреляции Спирмена

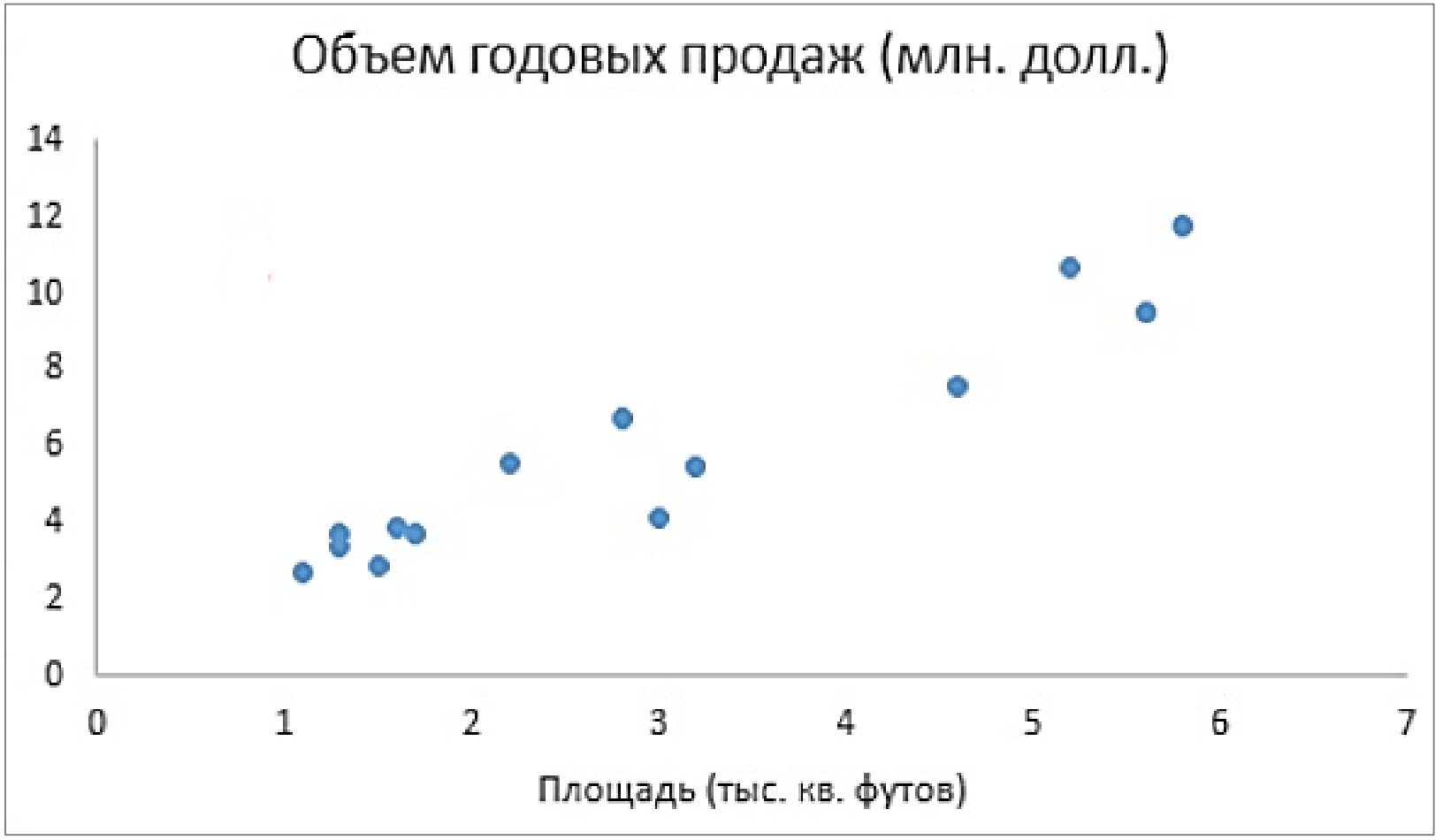


X	Y		X	Y	d^2
3,7	-0,3		1	1	0
5,8	4,1		2	4	4
7,1	4,3		3	5	4
9,1	8,3		4	7	9
10,5	12,9		5	10	25
11,4	8,1		6	6	0
11,6	9,5		7	9	4
12,5	9,5		8	8	0
14,3	18,2		9	12	9
17,5	17,5		10	11	1
30,0	1,0		11	2	81
32,0	2,0		12	3	81

$$r_s = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)}$$

Линейная регрессия

Годовые объемы продаж в 14 магазинах торговой сети.



Анализ диаграммы разброса показывает, что между площадью магазина *X* и годовым объемом продаж *Y* существует положительная зависимость. Если площадь магазина увеличивается, объем продаж возрастает почти линейно. Таким образом, в данном случае наиболее подходящей для исследования является линейная модель

Постановка задачи

Линейная регрессия некоторой зависимой переменной y на набор независимых переменных $x = (x_1, \ldots, x_r)$, где r – это число параметров объекта, предполагает, что между y и x линейное отношение:

$$y = \beta_o + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon.$$

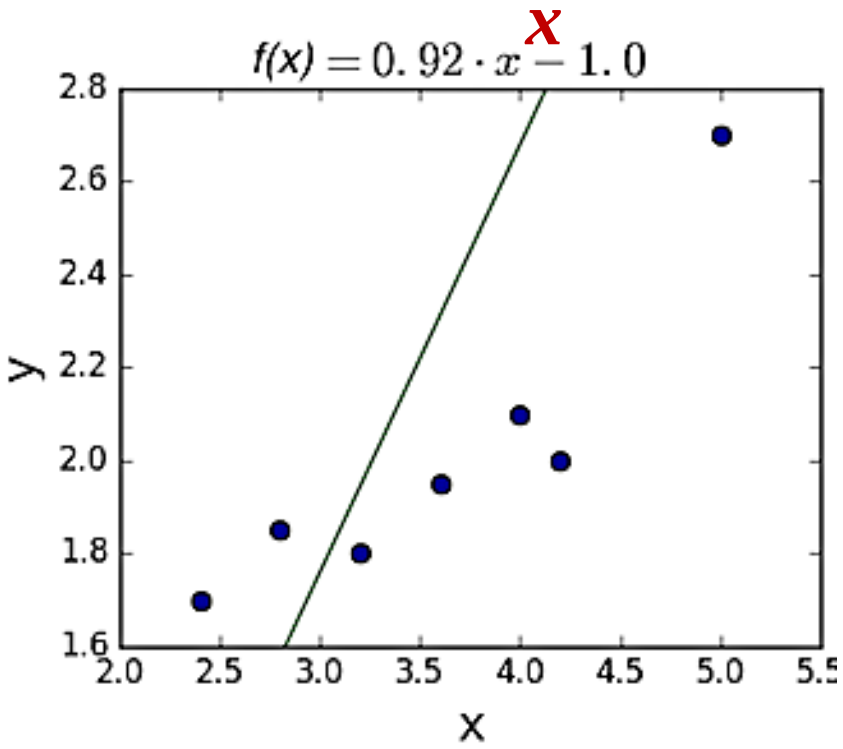
Это уравнение регрессии. β_o, β_1, …, β_r – коэффициенты регрессии, и ε – случайная ошибка.

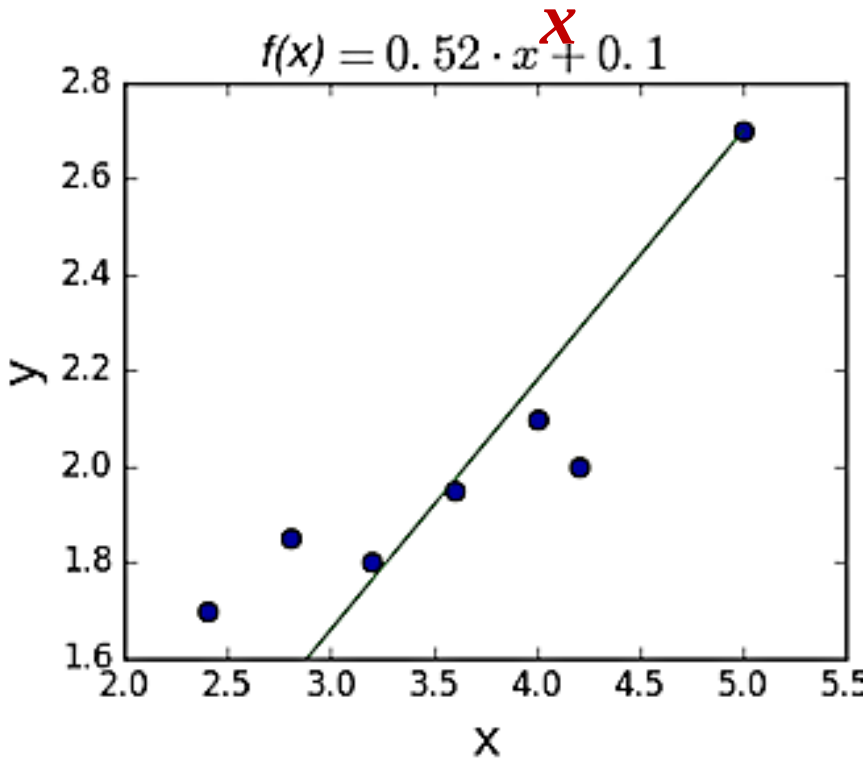
Линейная регрессия вычисляет коэффициенты регрессии или просто прогнозируемые веса измерения, обозначаемые как b_o, b_1, …, b_r. Они определяют оценочную функцию регрессии

$$f(x) = b_o + b_1 x_1 + \cdots + b_r x_r.$$

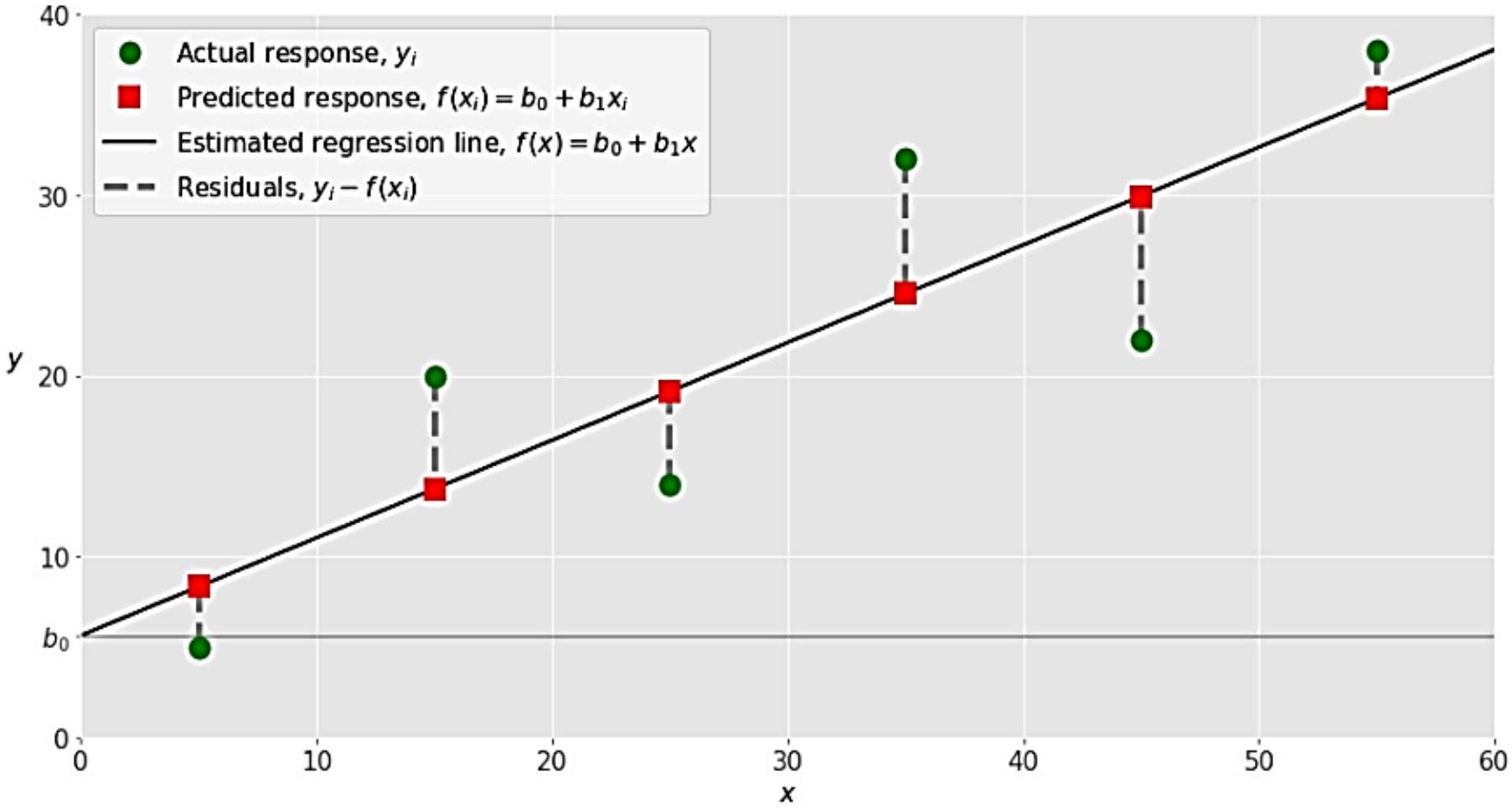
Простая линейная регрессия

Простая или одномерная линейная регрессия – случай линейной регрессии с единственной независимой переменной x.





Как выбрать правильную модель?



Для каждого результата наблюдения i = 1, …, n, оценочный или предсказанный ответ $f(x_i)$ должен быть как можно ближе к соответствующему фактическому ответу y_i. Разницы $y_i - f(x_i)$ для всех результатов наблюдений называются отклонениями.

Регрессия определяет лучшие прогнозируемые веса измерения, которые соответствуют наименьшим отклонениям.

Функция линейной регрессии выражается уравнением

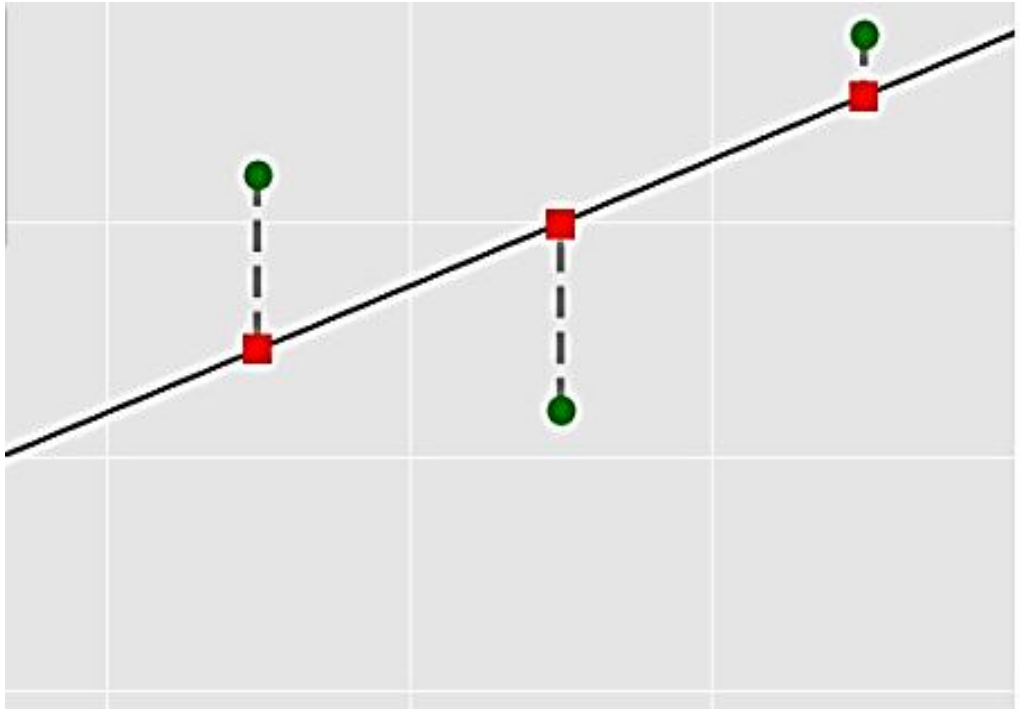
$$f(x) = b_0 + b_1 x$$

Для определения функции регрессии необходимо рассчитать оптимальные значения коэффициентов b_0 и b_1 для минимизации отклонений фактических значений зависимой величины от предсказанных.

Для получения лучших весов, нужно минимизировать сумму квадратов отклонений (SSR) для всех результатов наблюдений:

$$SSR = \Sigma_i (y_i - f(x_i))^2$$

Этот подход называется **методом наименьших квадратов**.

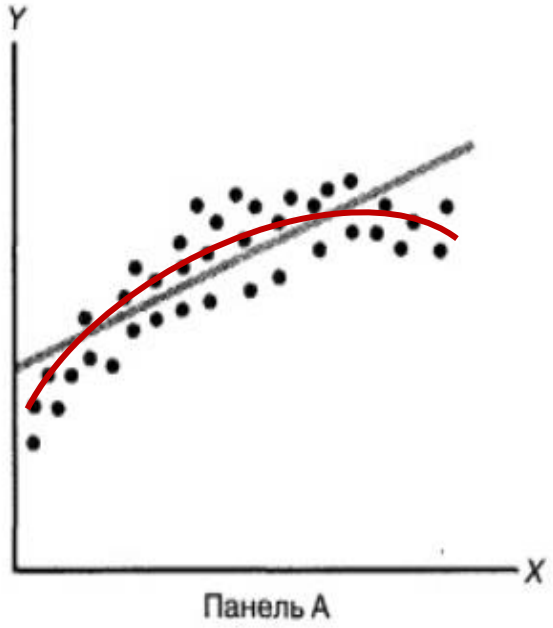


Ограничения линейной регрессии

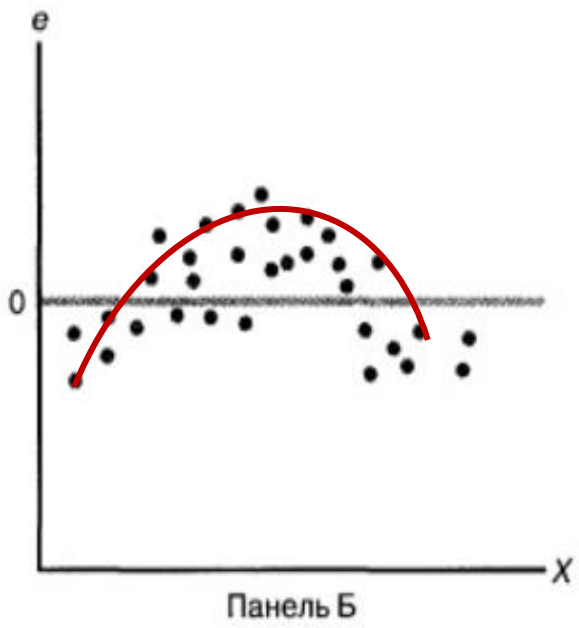
https://habr.com/ru/post/350668/

Для того, чтобы корректно использовать модель линейной регрессии необходимы некоторые допущения относительно распределения и свойств переменных.

1. Должно соблюдаться условие линейности
2. Ошибка должна иметь нормальное распределение.
3. Вариация данных вокруг линии регрессии должна быть постоянной.
4. Ошибки должны быть независимыми.



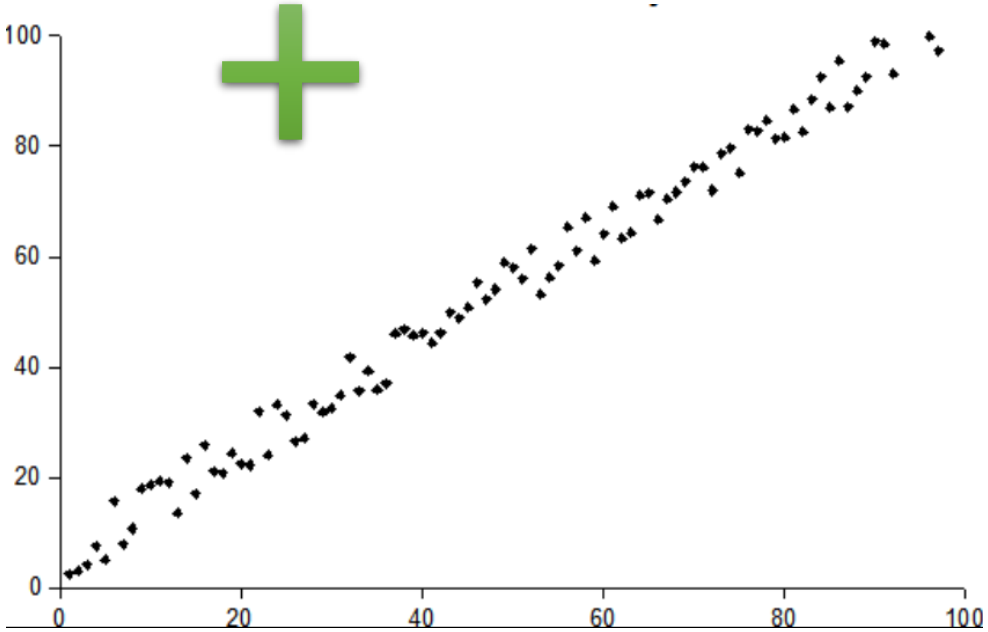
Панель А

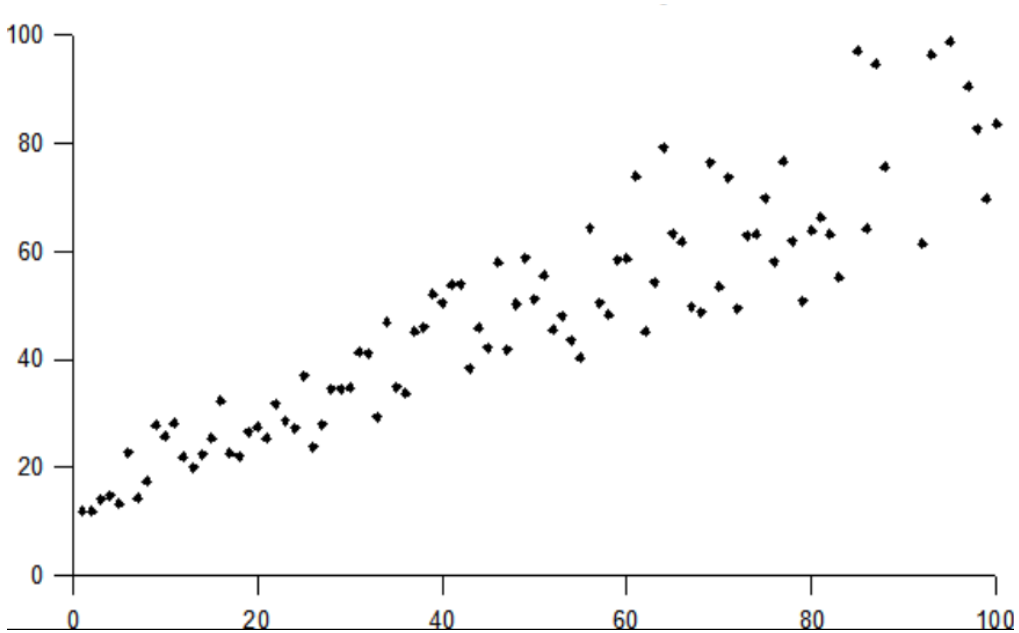


Панель Б

Панель А иллюстрирует возрастание переменной *Y* при увеличении переменной *X*. Однако зависимость между этими переменными носит нелинейный характер, поскольку скорость возрастания переменной *Y* падает при увеличении переменной *X*. Таким образом, для аппроксимации зависимости между этими переменными лучше подойдет квадратичная модель. Особенно ярко квадратичная зависимость между величинами X_i и e_i проявляется на панели Б.

Условие 3: Вариация данных вокруг линии регрессии должна быть постоянной





Функция регрессии выражается уравнением

$$f(x) = b_o + b_1x$$

Величина b_o, также называемая сдвигом, показывает точку, где расчётная линия регрессии пересекает ось y. Это значение расчётного ответа f(x) для x = 0. Величина b_1 определяет наклон расчетной линии регрессии.

Пример 1. Один экономист решил предсказать изменение индекса 500 наиболее активно покупаемых акций на Нью-Йоркской фондовой бирже, публикуемого агентством Standard and Poor, на основе показателей экономики США за 50 лет. В результате он получил следующее уравнение линейной регрессии: $\hat{Y}i = -5,0 + 7Xi$. Какой смысл имеют параметры сдвига b0 и наклона b1.

Решение. *Сдвиг регрессии b0 равен –5,0. Это значит, что, если рост экономики США равен нулю, индекс акций за год снизится на 5%. Наклон b1 равен 7. Следовательно, при увеличении темпов роста экономики на 1% индекс акций возрастает на 7%.*

Пример множественной регрессии по прогнозированию цены дома в Бостоне

Взяли два параметра, получили два коэф. или два веса для параметров

```
array([-0.34977589, 0.11642402])
```

Например, первый признак нашего решения — это количество преступлений на одного жителя. Коэффициент этого признака нашей модели составил –0.35. Это значит, что если мы умножим этот коэффициент на 1000 (т. к. вектором целей является цена дома в тысячах долларов), то у нас будет изменение в цене дома для каждого дополнительного преступления на душу населения:

```
# Первый коэффициент, умноженный на 1000
model.coef_[0]*1000
```

```
-349.77588707748947
```

Это говорит о том, что каждое преступление на душу населения снизит цену дома примерно на $350!

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r^2 , называемый коэффициентом детерминации.

Коэффициент детерминации характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака:

$$r_{xy}^2 = \frac{\sigma_{y_{объясн.}}^2}{\sigma_{y_{общ.}}^2}$$

Коэффициент детерминации изменяется в диапазоне от $-\infty$ до 1.

Если он равен 1, это соответствует идеальной модели, когда все точки наблюдений лежат точно на линии регрессии, т.е. сумма квадратов их отклонений равна 0.

Если коэффициент детерминации равен 0, это означает, что связь между переменными регрессионной модели отсутствует, и вместо нее для оценки значения выходной переменной можно использовать простое среднее ее наблюдаемых значений.

Например = 0,982. Таким образом, уравнением регрессии объясняется 98,2% дисперсии выходной переменной , а на долю прочих факторов приходится лишь 1,8% ее дисперсии (т.е. остаточная дисперсия). Величина коэффициента детерминации является одним из критериев оценки качества линейной модели.

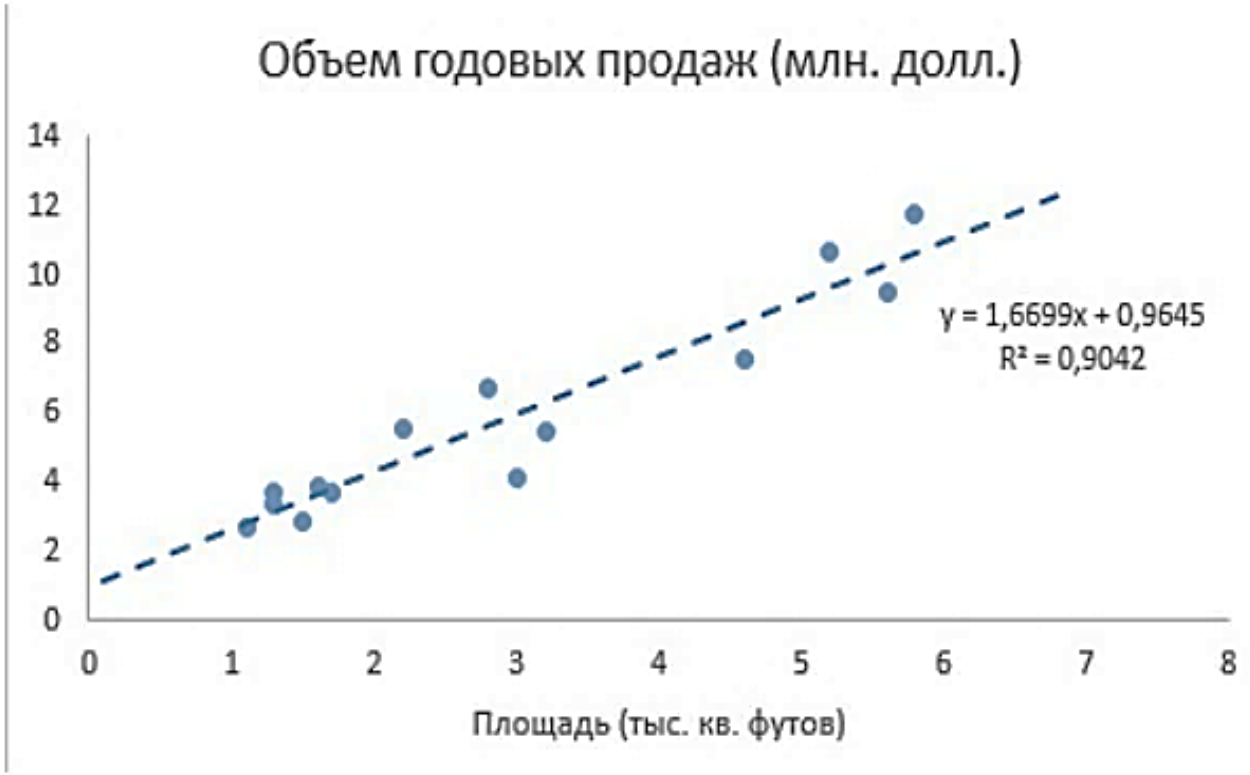
Прогнозирование в регрессионном анализе: **интерполяция и экстраполяция**

Интерполяция – предсказывание значения Y между значениями переменной X в диапазоне возможных значений.

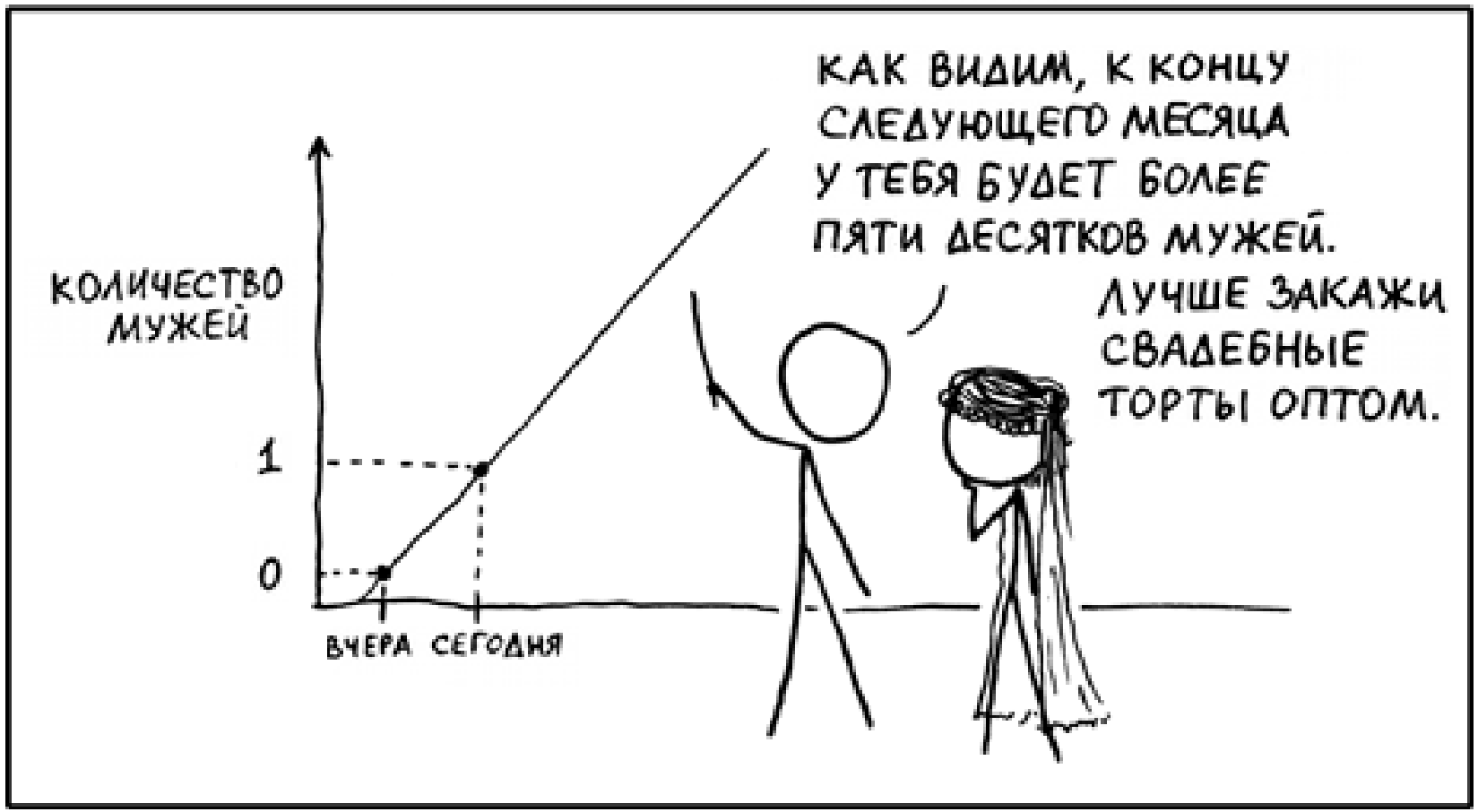
Экстраполяция – предсказывание значения Y за пределами этого интервала

Экстраполяция не всегда релевантна.
Например, пытаясь предсказать среднегодовой объем продаж в магазине, зная его площадь, можно вычислять значение переменной Y лишь для значений X от 1,1 до 5,8 тыс. кв. футов.

Любая попытка экстраполяции означает, что мы предполагаем, будто линейная регрессия сохраняет свой характер за пределами допустимого диапазона.



МОЁ ХОББИ: ЭКСТРАПОЛИРОВАТЬ



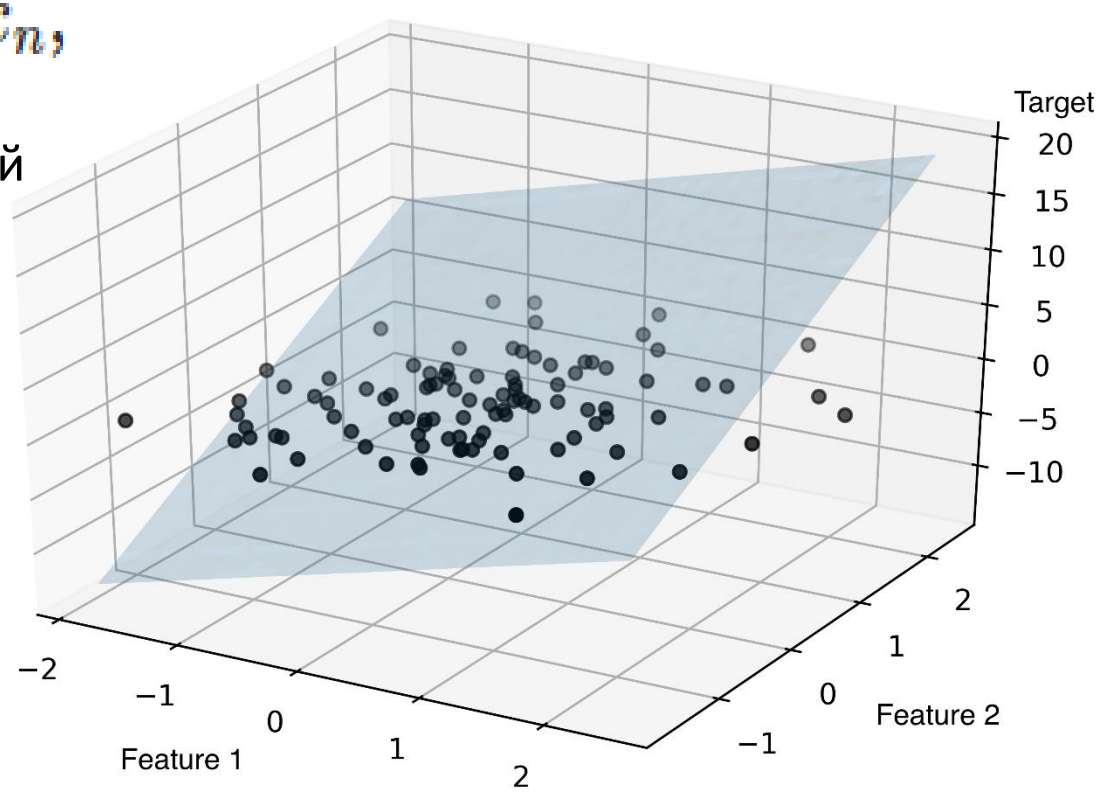
Множественная линейная регрессия

Множественной называют линейную регрессию, в модели которой число независимых переменных две или более.
Уравнение множественной линейной регрессии имеет вид:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n,$$

Отличие между простой и множественной линейной регрессией заключается в том, что вместо линии регрессии в ней используется гиперплоскость.

Использование в модели нескольких входных переменных позволяет увеличить долю объяснённой дисперсии выходной переменной, и таким образом улучшить соответствие модели данным. Т.е. **при добавлении в модель каждой новой переменной коэффициент детерминации растёт**.



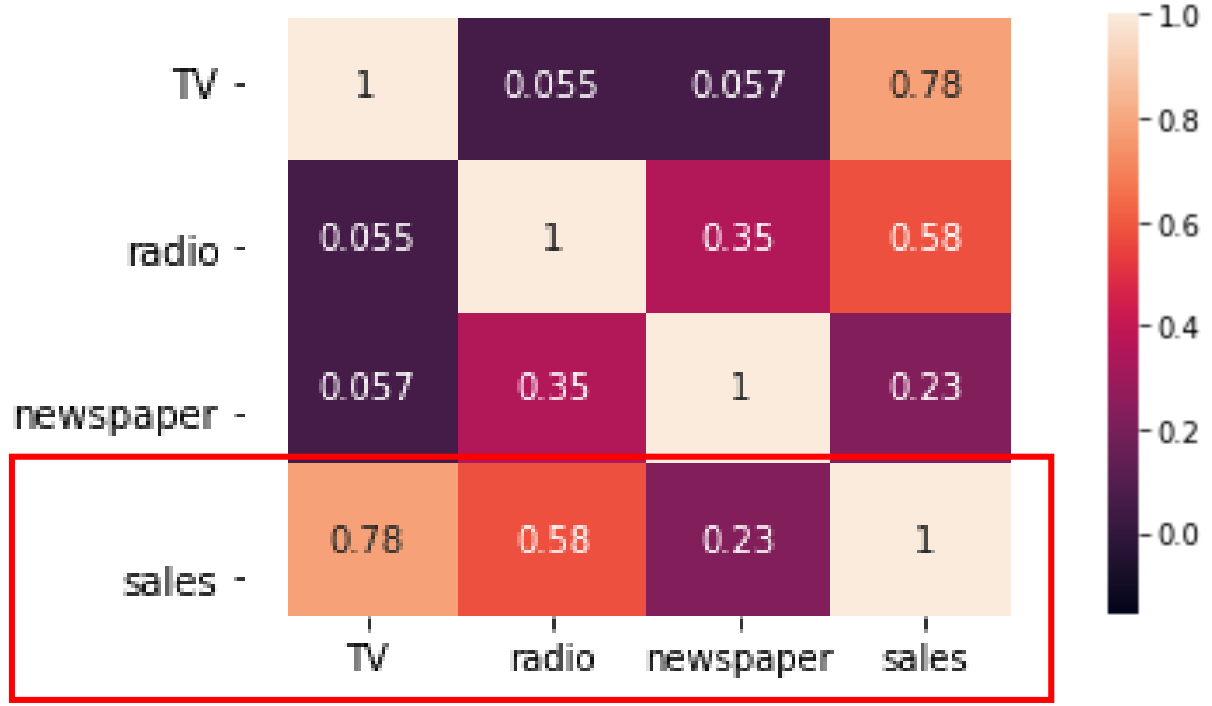
Пример: Датасет зависимости доходов на распродажах от рекламы

```
data = pd.read_csv("data/reklama.csv")
```

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

Рассчитаем коэффициенты корреляции между параметрами и визуализируем их при помощи тепловой карты

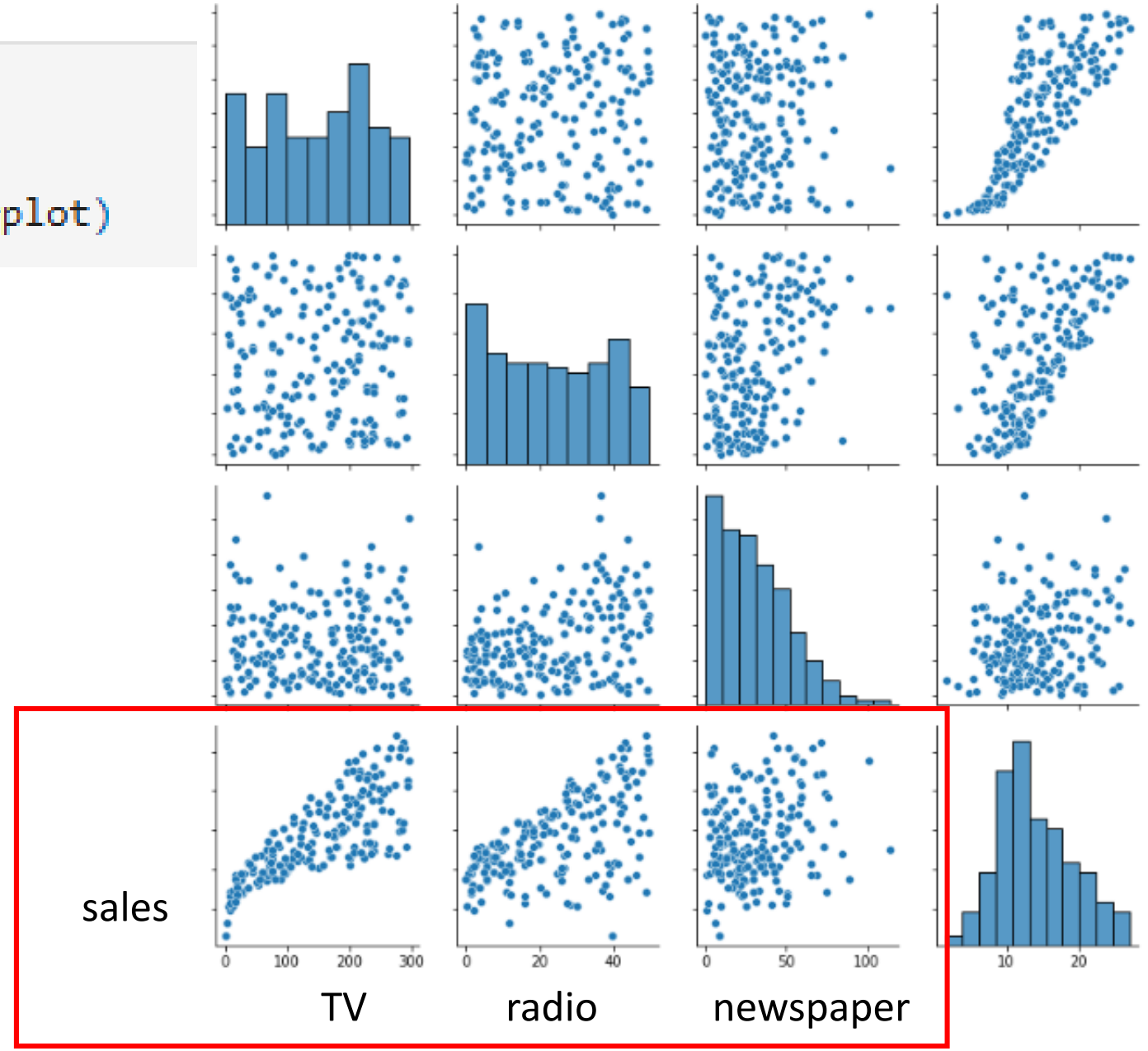
```
hm = sns.heatmap(data.corr(),
                 cbar=True,
                 annot=True)
```



Из матрицы корреляции видно, что самый большой коэффициент корреляции между продажами и рекламой на ТВ.

Построим матрицу диаграмм разброса, чтобы оценить связи визуально.

```
g = sns.PairGrid(data)
g.map_diag(sns.histplot)
g.map_offdiag(sns.scatterplot)
```

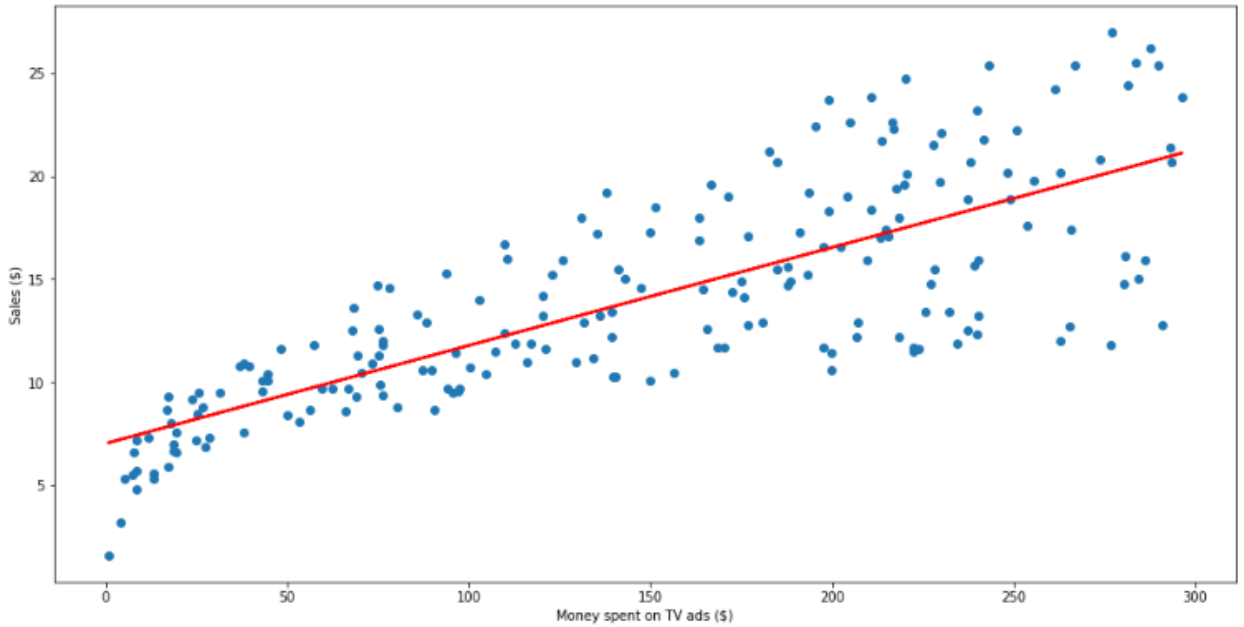


Получим модель линейной регрессии зависимости продаж от рекламы на ТВ

```
X = data['TV'].values.reshape(-1,1)
y = data['sales'].values.reshape(-1,1)

reg = LinearRegression()
reg.fit(X, y)

print(reg.coef_[0][0])
print(reg.intercept_[0])
print("The linear model is: Y = {:.5} + {:.5}*TV".format(reg.intercept_[0],
reg.coef_[0][0]))
```

```
0.04753664043301975
7.032593549127695

The linear model is:

Y = 7.0326 + 0.047537*TV
```

Визуализируем полученную модель на диаграмме разброса



Получим модель множественной линейной регрессии для оценки вклада в продажи остальных параметров выборки

```
Xs = data.drop(['sales', 'Unnamed: 0'], axis=1)
y = data['sales'].values.reshape(-1,1)

reg = LinearRegression()
reg.fit(Xs, y)
LinearRegression()
print(reg.coef_)
print(reg.intercept_)
[[ 0.04576465  0.18853002 -0.00103749]]
[2.93888937]

print("The linear model is: Y = {:.5} + {:.5}*TV + {:.5}*radio +
{:.5}*newspaper".format(reg.intercept_[0], reg.coef_[0][0], reg.coef_[0][1],
reg.coef_[0][2]))

The linear model is: Y = 2.9389 + 0.045765*TV + 0.18853*radio + -0.0010375*newspaper
```

Рассчитаем коэффициент детерминации для обоих моделей из примера

- ✓ Коэффициент детерминации модели линейной регрессии зависимости продаж от рекламы на ТВ

0.611875050850071

- ✓ Коэффициент детерминации модели множественной линейной регрессии зависимости продаж от всех видов рекламы

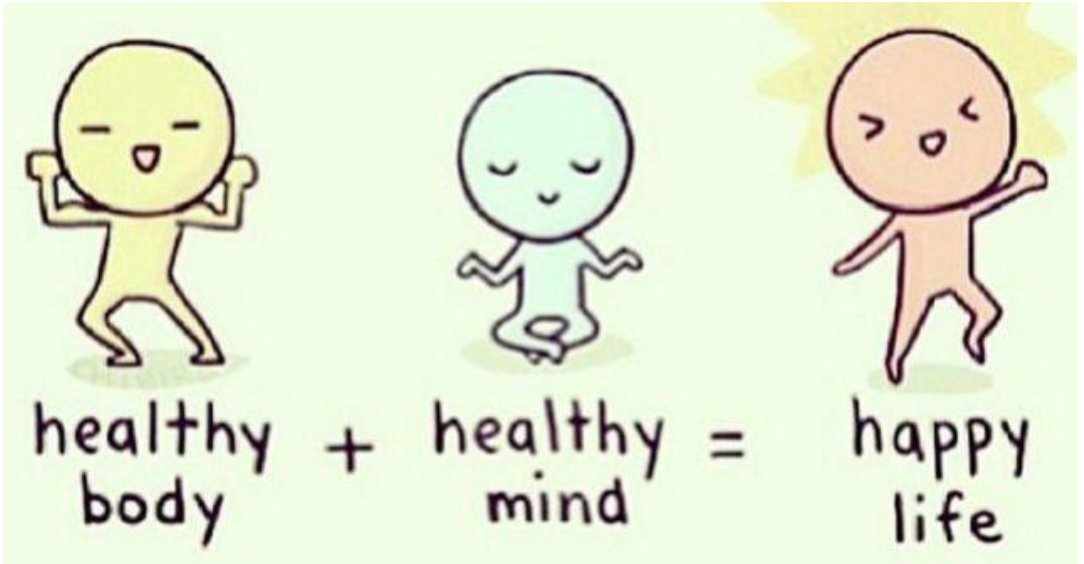
0.8972106381789522

Коэффициент детерминации измеряет степень соответствия модели реальным данным. Он показывает, какую долю вариаций зависимой переменной можно объяснить независимой переменной или переменными.

https://www.kaggle.com/code/hely333/eda-regression/notebook

Пример: Рассмотрим **набор данных, посвященный расходам на лечение** разных пациентов.

Нас интересует какие факторы в наибольшей степени влияют на стоимость лечения



	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Предобработка данных

1

```
data.isnull().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

2

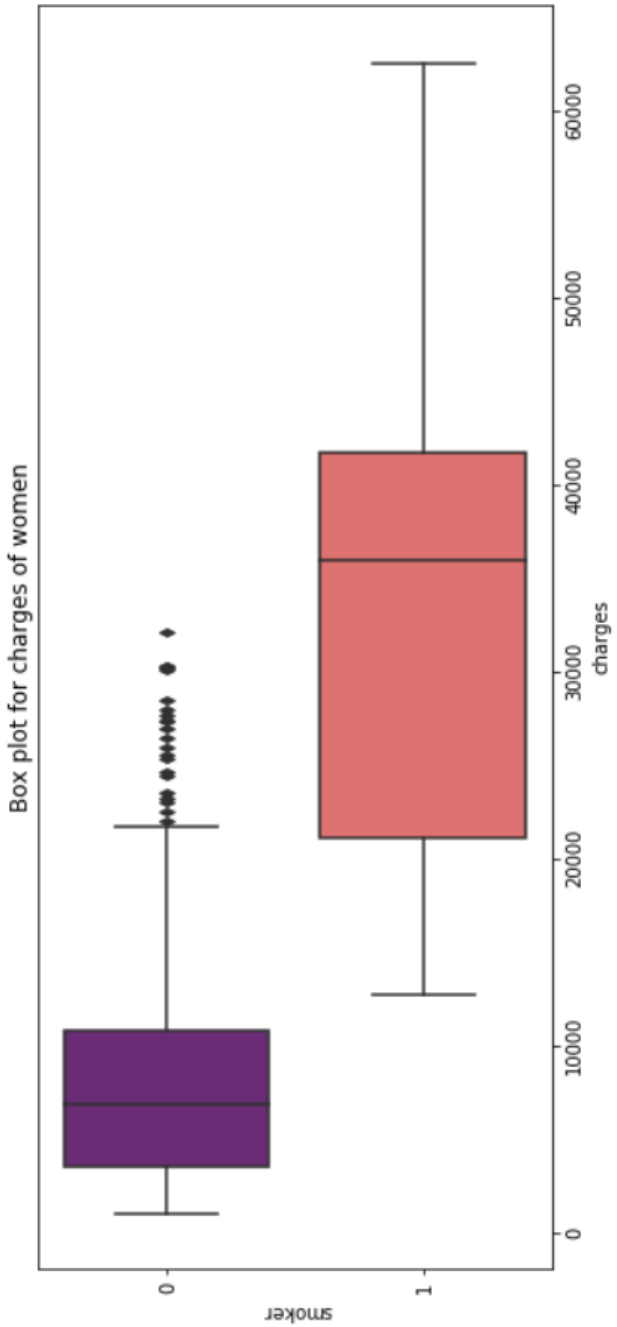
```
from sklearn.preprocessing import LabelEncoder
#sex
le = LabelEncoder()
le.fit(data.sex.drop_duplicates())
data.sex = le.transform(data.sex)
# smoker or not
le.fit(data.smoker.drop_duplicates())
data.smoker = le.transform(data.smoker)
#region
le.fit(data.region.drop_duplicates())
data.region = le.transform(data.region)
```

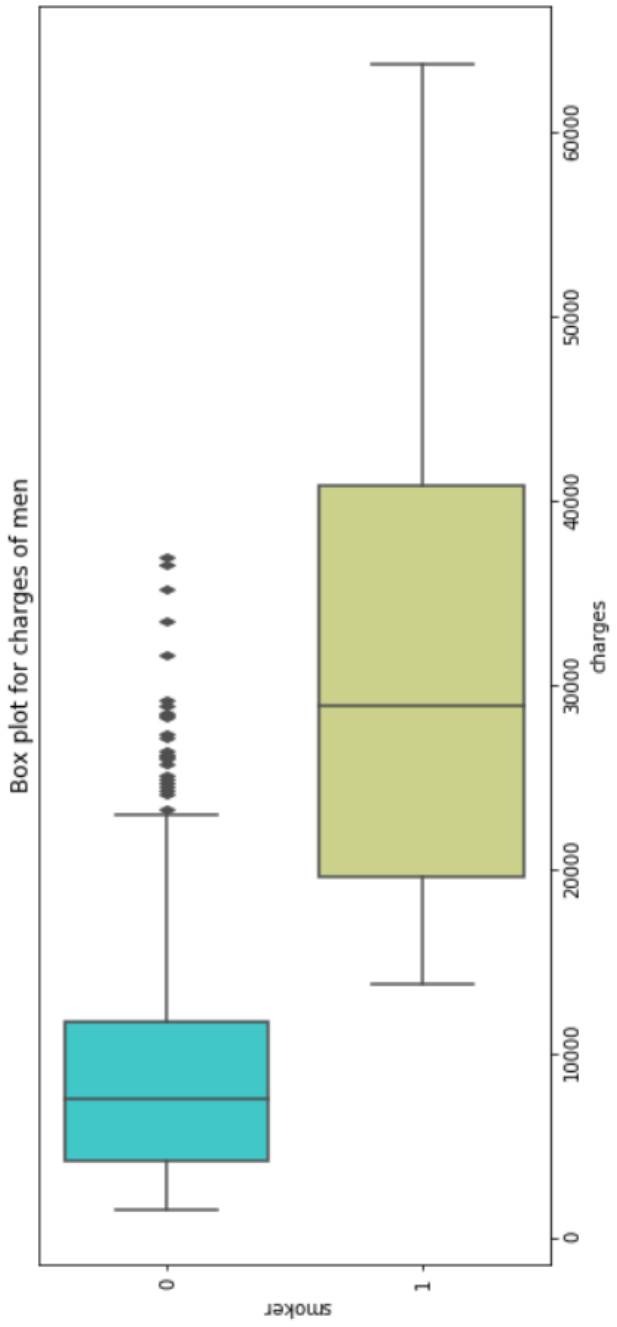
```
data.corr()['charges'].sort_values()
```

```
region     -0.006208
sex         0.057292
children    0.067998
bmi         0.198341
age         0.299008
smoker      0.787251
charges     1.000000
Name: charges, dtype: float64
```

Рассчитаем **коэффициенты корреляции** различных факторов с расходами на лечение.
И видим, что **самая большая корреляция расходов на лечение с показателем курение пациента**

Женщины

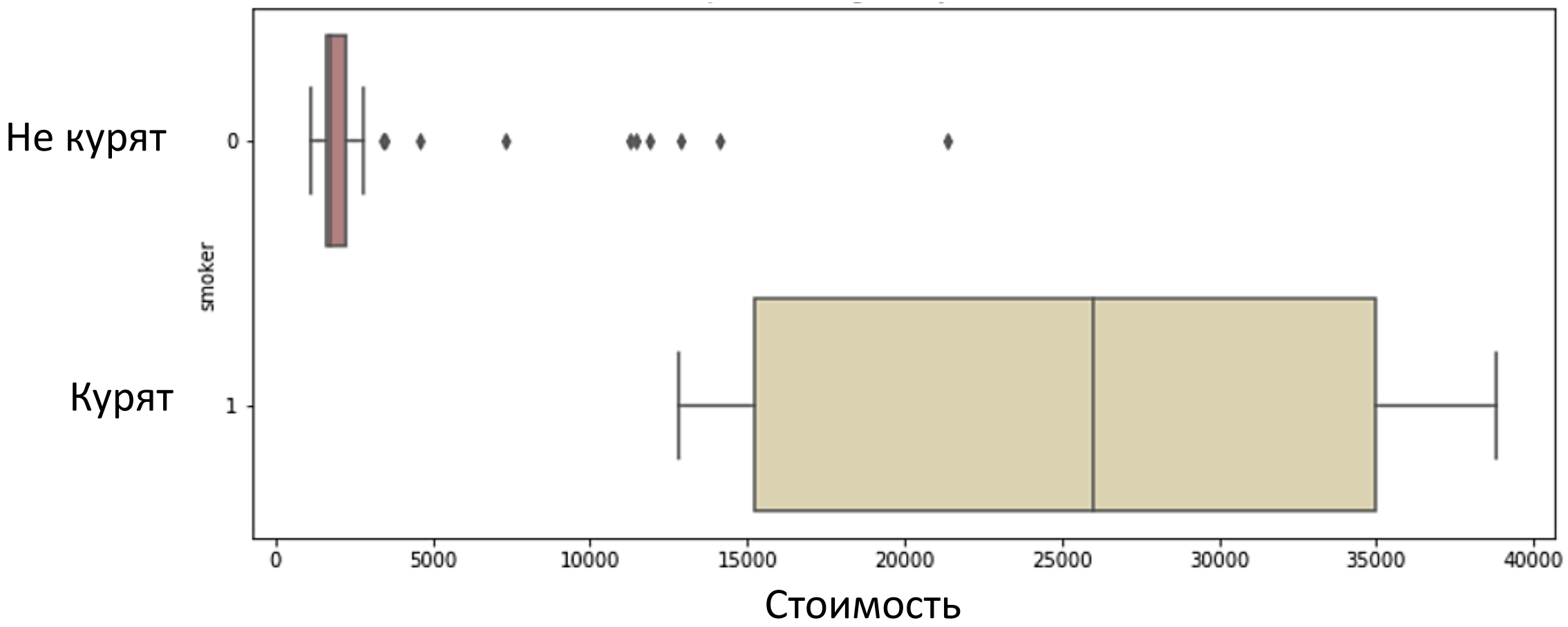




Мужчины

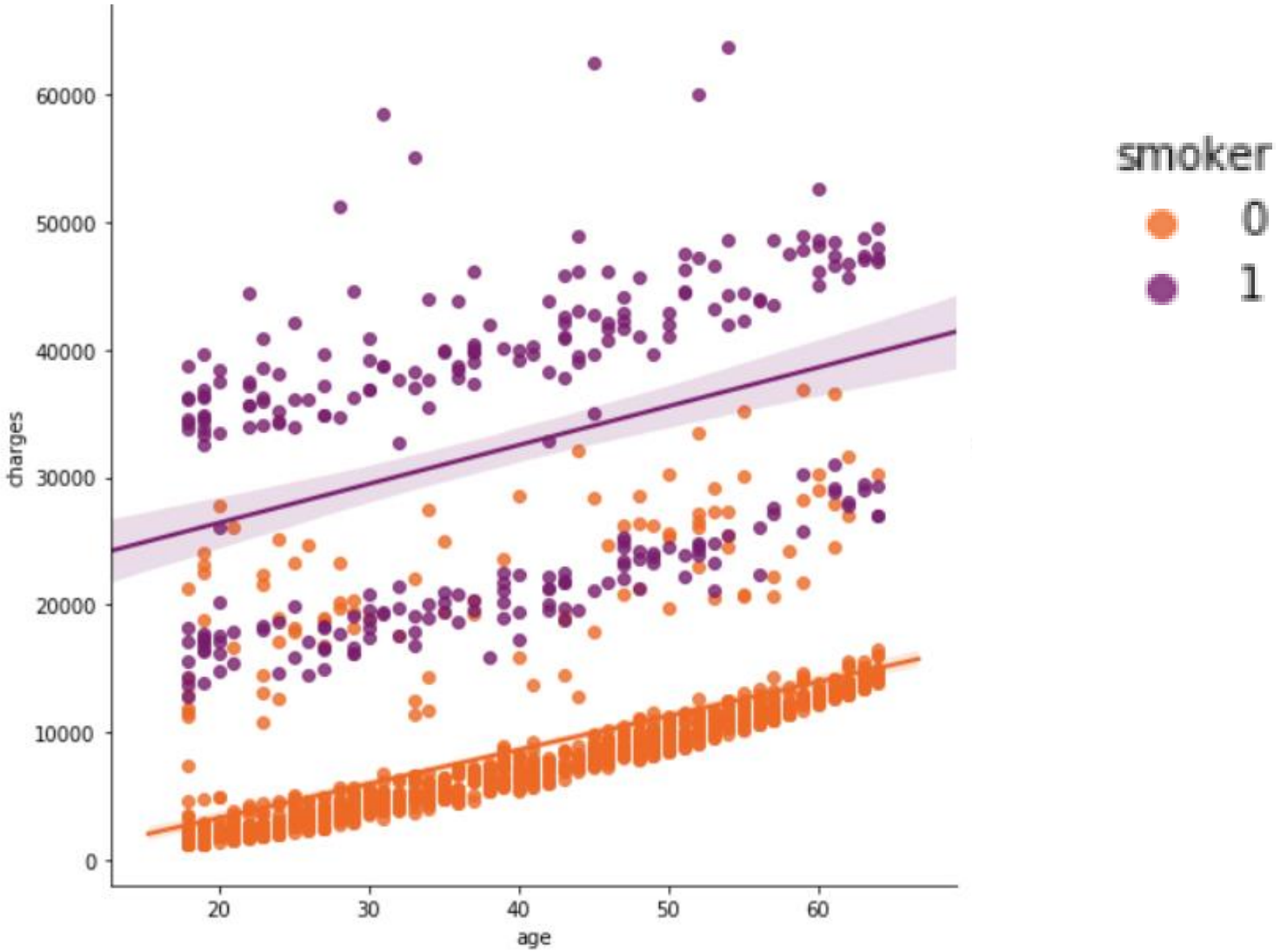
Стоимость лечения у женщин и мужчин примерно одинакова

Box plot for charges 18 years old smokers

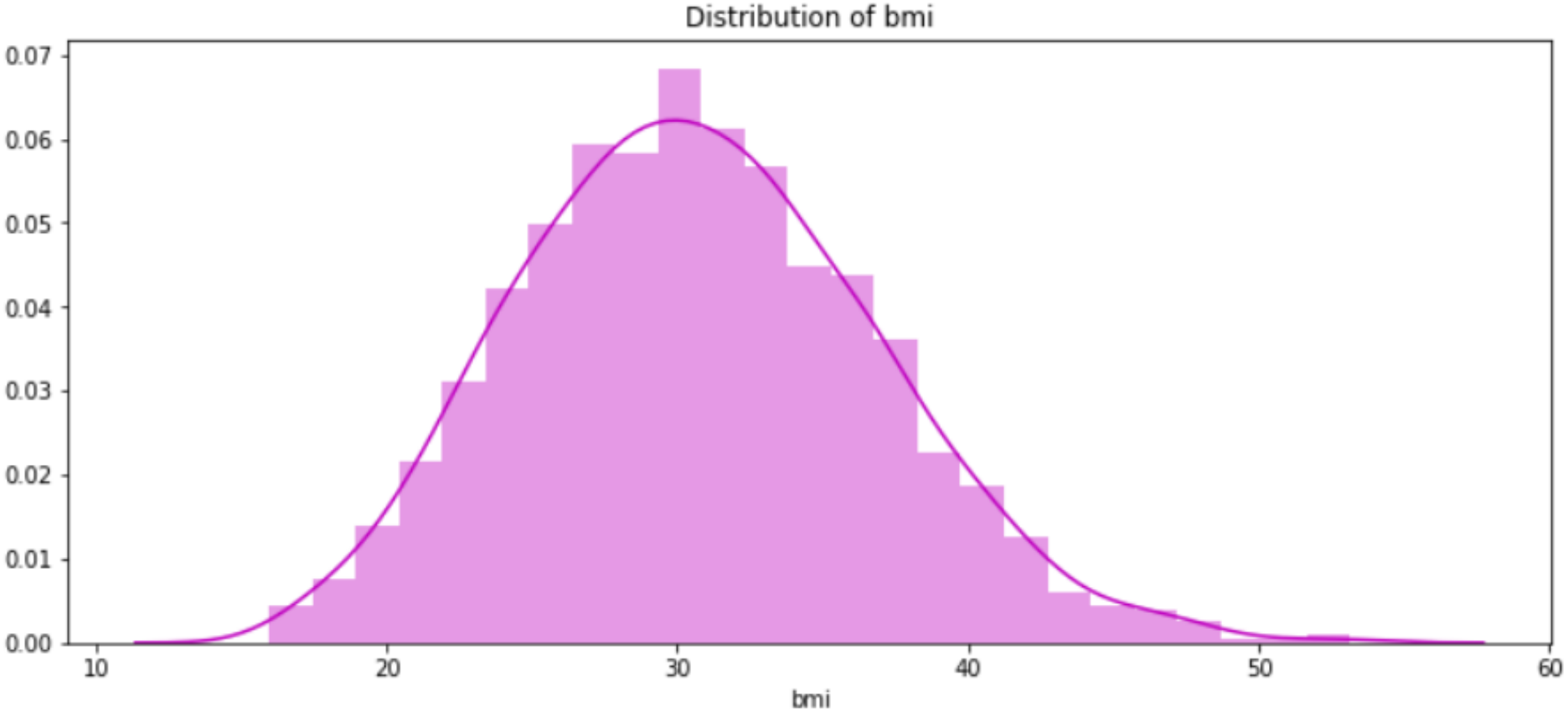


Распределение стоимости лечения курящих и некурящих 18-летних пациентов ! ! !

Распределение расходов на лечение с возрастом

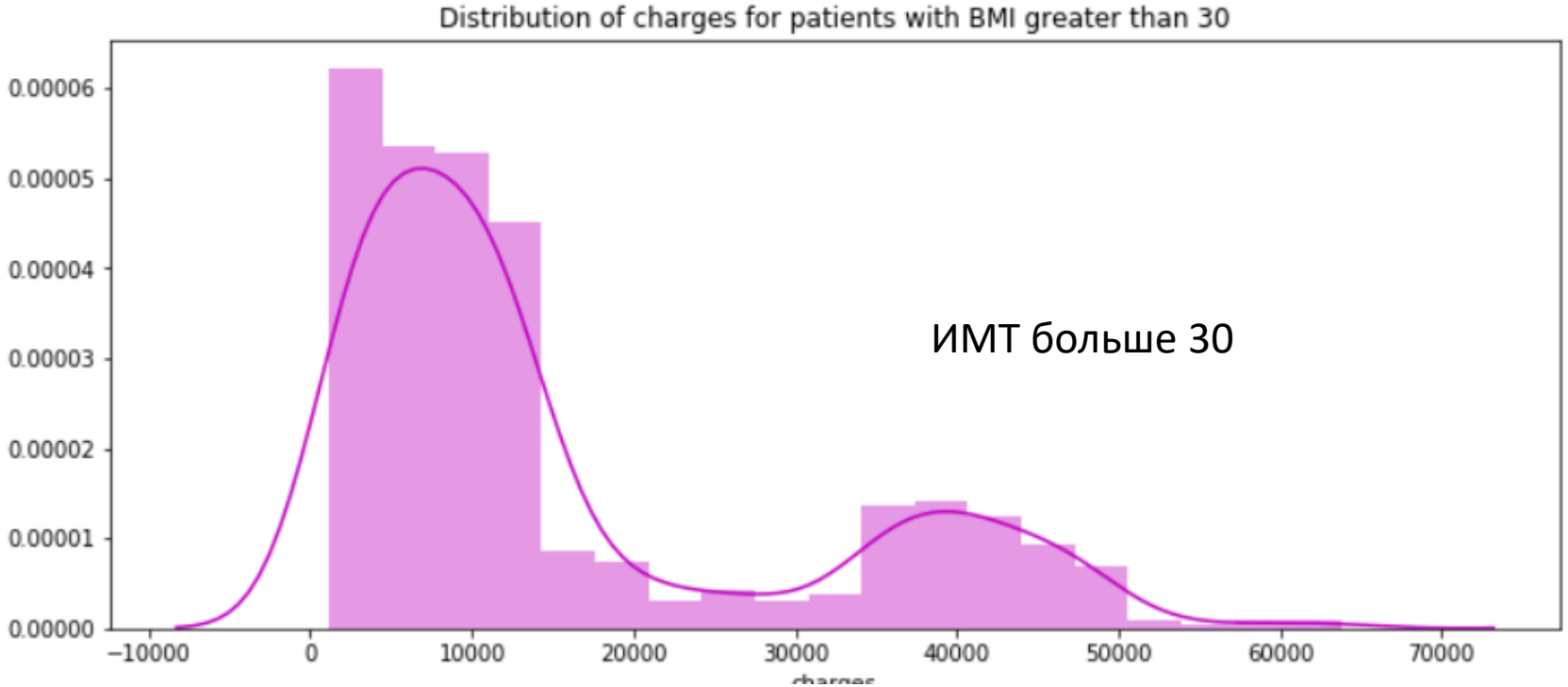


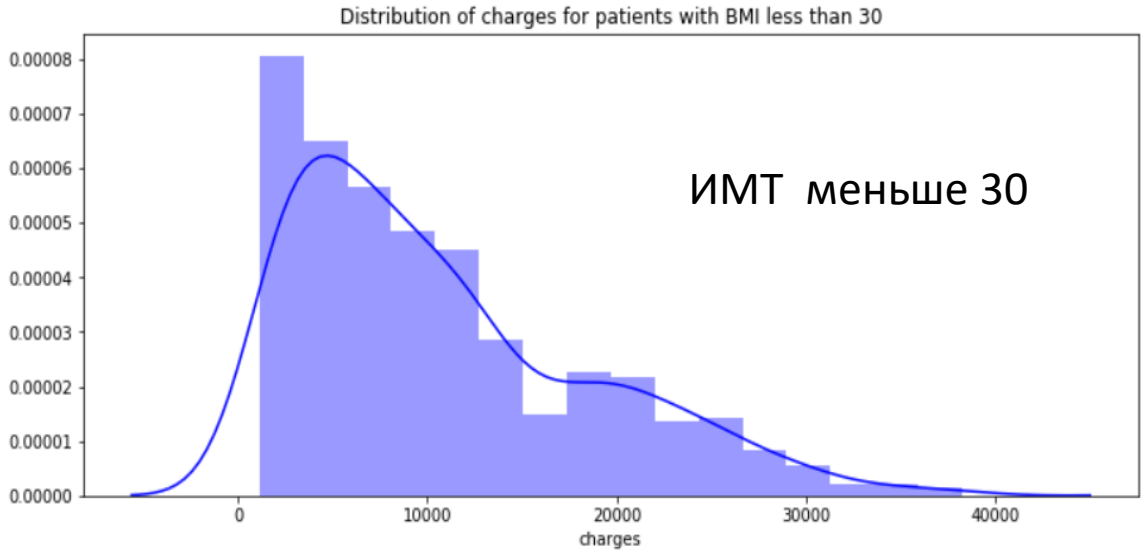
Давайте обратим внимание на ИМТ (индекс массы тела).



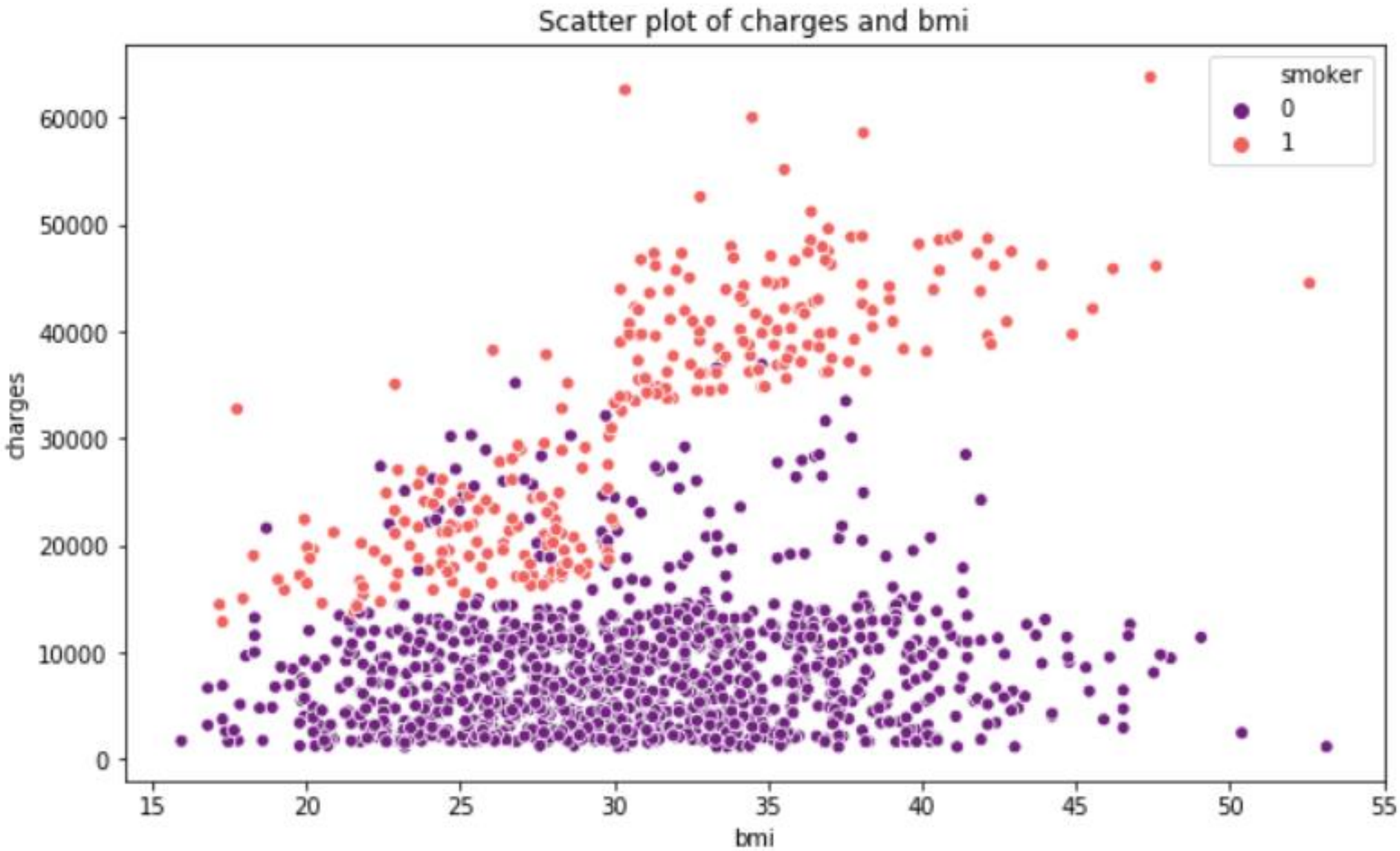
При значении, равном 30, начинается ожирение.

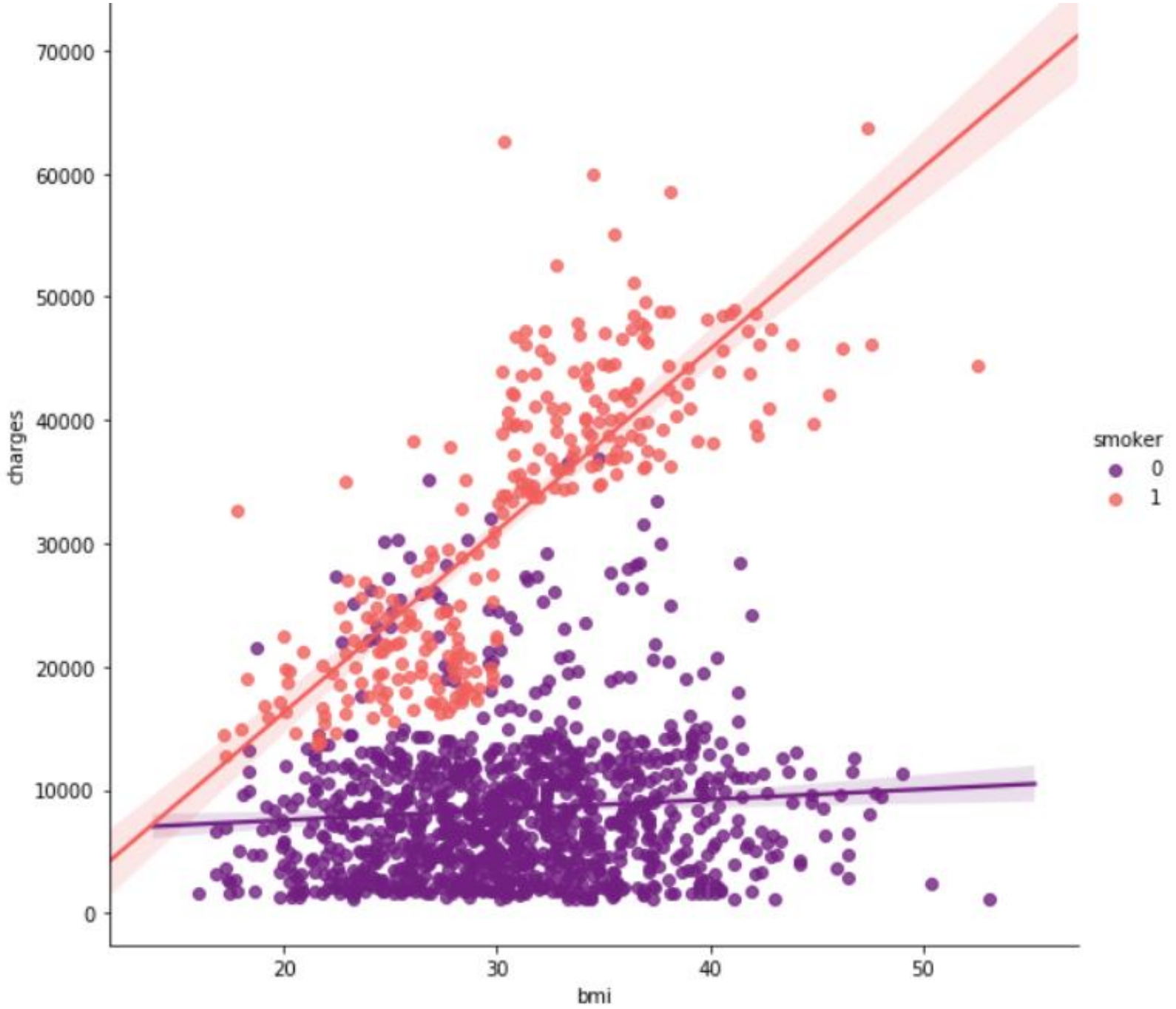
Для начала давайте рассмотрим **распределение расходов** у пациентов с ИМТ больше 30 и меньше 30.





Для пациентов с ожирением расходы превысили доходят до 70тыс, в то время как для остальных не превышают 40тыс


Scatter plot of charges and bmi
smoker
0
1
charges
bmi
0
10000
20000
30000
40000
50000
60000
15
20
25
30
35
40
45
50
55



Здесь визуализированы модели линейных регрессий зависимости стоимости лечения от индекса массы тела и курения пациентов.

Делайте выводы :)