Лекция 6 Регрессионный анализ

- Корреляция
- Коэффициент корреляции
- Линейная регрессия
- Множественная линейная регрессия

Рассмотрим методы предсказания значений числовой переменной в зависимости от значений одной или нескольких других числовых переменных.

Как правило, для предсказания значений переменной используется **регрессионный анализ.** Его цель — разработать статистическую модель, позволяющую предсказывать значения зависимой переменной (переменной отклика) по значениям независимых (объясняющих) переменных. Если независимых переменных более одной, то речь идет о множественной регрессии, в случае одной независимой переменной будет рассматриваться линейная регрессия.

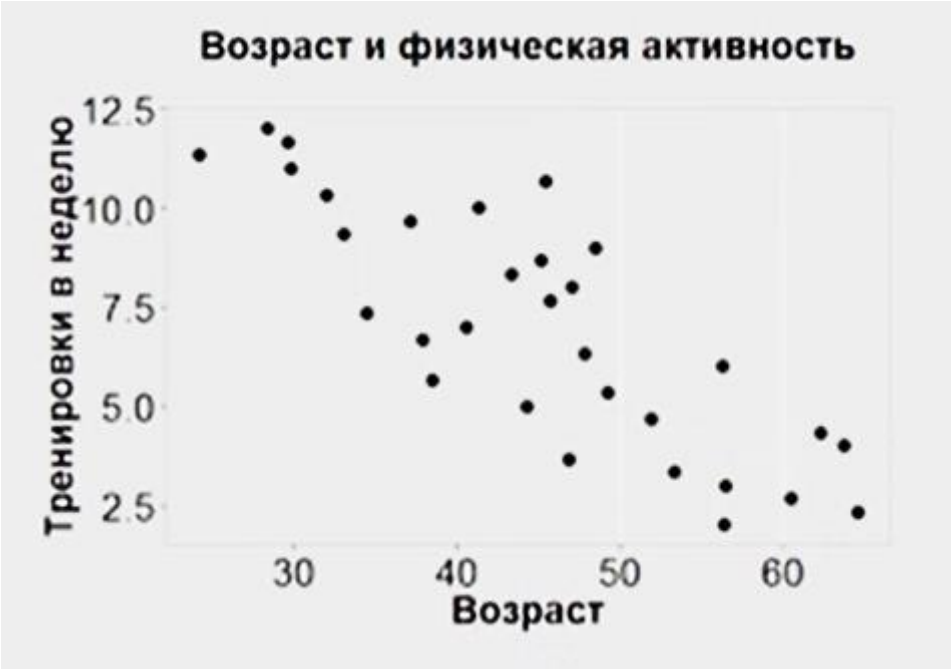
Корреляция

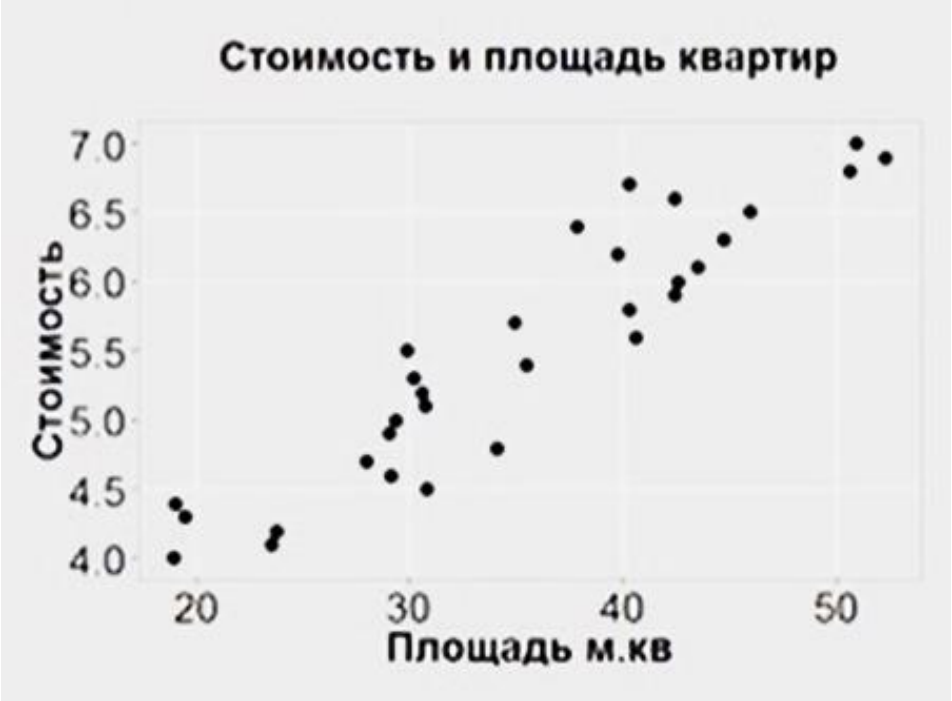
Корреля́ция (от лат. correlatio «соотношение»)— статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению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

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

Пример: существует корреляция между температурой воздуха и потреблением мороженого. Чем жарче погода, тем больше мороженого покупают.

В качестве примеров рассмотрим диаграммы рассеивания, где по осям отложены два параметра, взаимосвязь которых мы хотим оценить:





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, рассмотрим пример прямой корреляции: чем выше уровень благосостояния человека, тем больше его продолжительность жизни. Обеспеченные люди питаются качественной пищей и своевременно получают врачебную помощь. В отличие от бедняков. Однако нельзя с уверенностью сказать, что определенный олигарх проживет дольше вот этого 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.

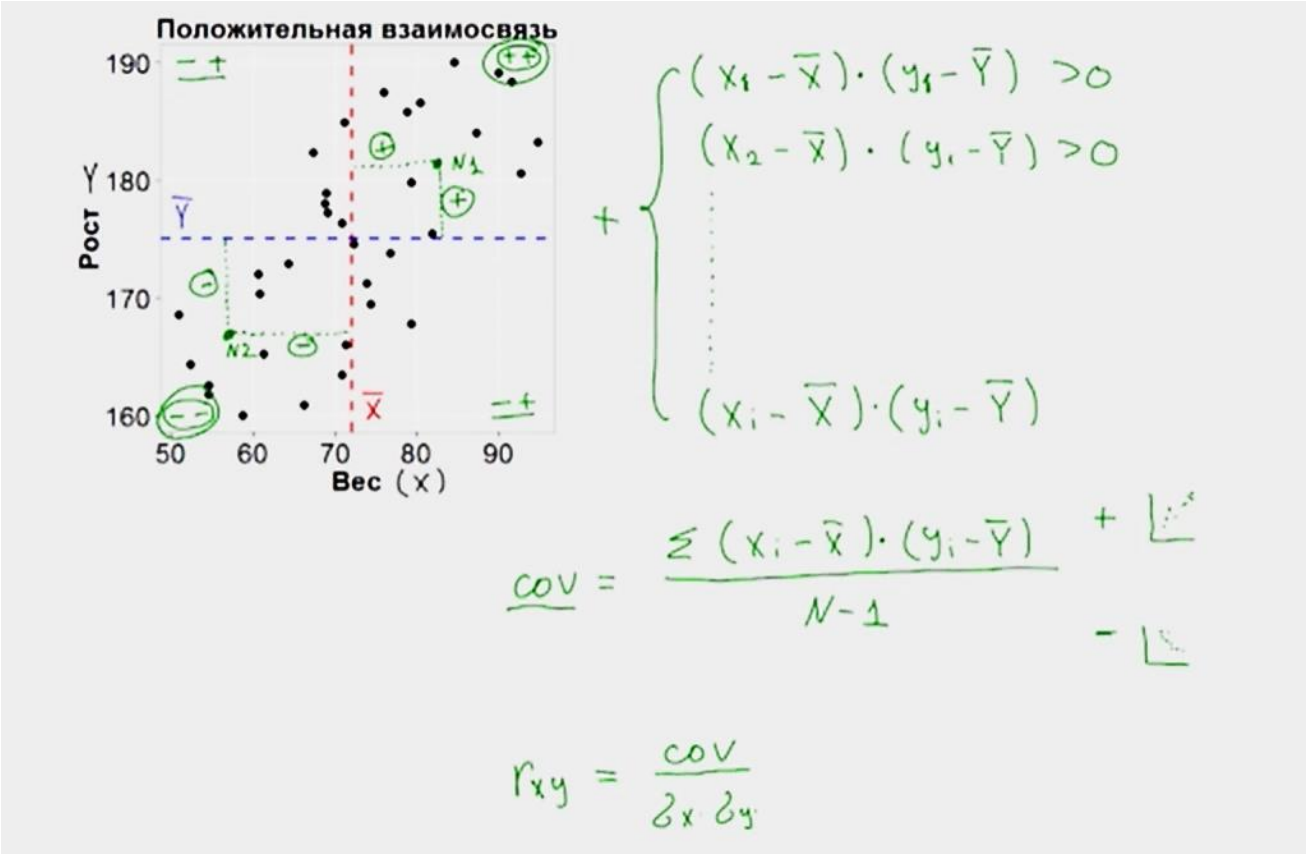
Коэффициент корреляции

Математической мерой корреляции двух случайных величин служит **коэффициент корреляции**. Он отражает силу и направление взаимосвязи величин и находится в промежутке от -1 до 1.

Посмотрим на примере:

Значение коэффициента	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!

Как рассчитать величину коэффициента корреляции:



https://stepik.org/lesson/8086/step/4?unit=1365

cov – ковариация

$\bar{x}$ - среднее значение по выборке

$\sigma_x \sigma_y$ - стандартные отклонения

Давайте остановимся на формуле коэффициента корреляции, которую мы получили:

$$r_{xy} = \frac{cov_{xy}}{\sigma_x \sigma_y}$$

запишем формулу чуть подробнее и выполним возможные преобразования:

$$r_{xy} = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{(N-1)\sigma_x \sigma_y} =$$

$$\frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{(N-1)\sqrt{\frac{\sum (x_i - \bar{X})^2}{N-1}}\sqrt{\frac{\sum (y_i - \bar{Y})^2}{N-1}}} =$$ # распишем подробнее стандартные отклонения

$$\frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{(N-1)\frac{1}{(N-1)}\sqrt{\sum (x_i - \bar{X})^2}\sqrt{\sum (y_i - \bar{Y})^2}} =$$ # теперь вынесем 1/ (N - 1) из под корней

$$\frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2}\sqrt{\sum (y_i - \bar{Y})^2}}$$ # и сократим (N - 1)

таким образом, мы сократили N - 1 в знаменателе и получили финальную формулу для коэффициента корреляции, которую вы часто сможете встретить в учебниках:

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

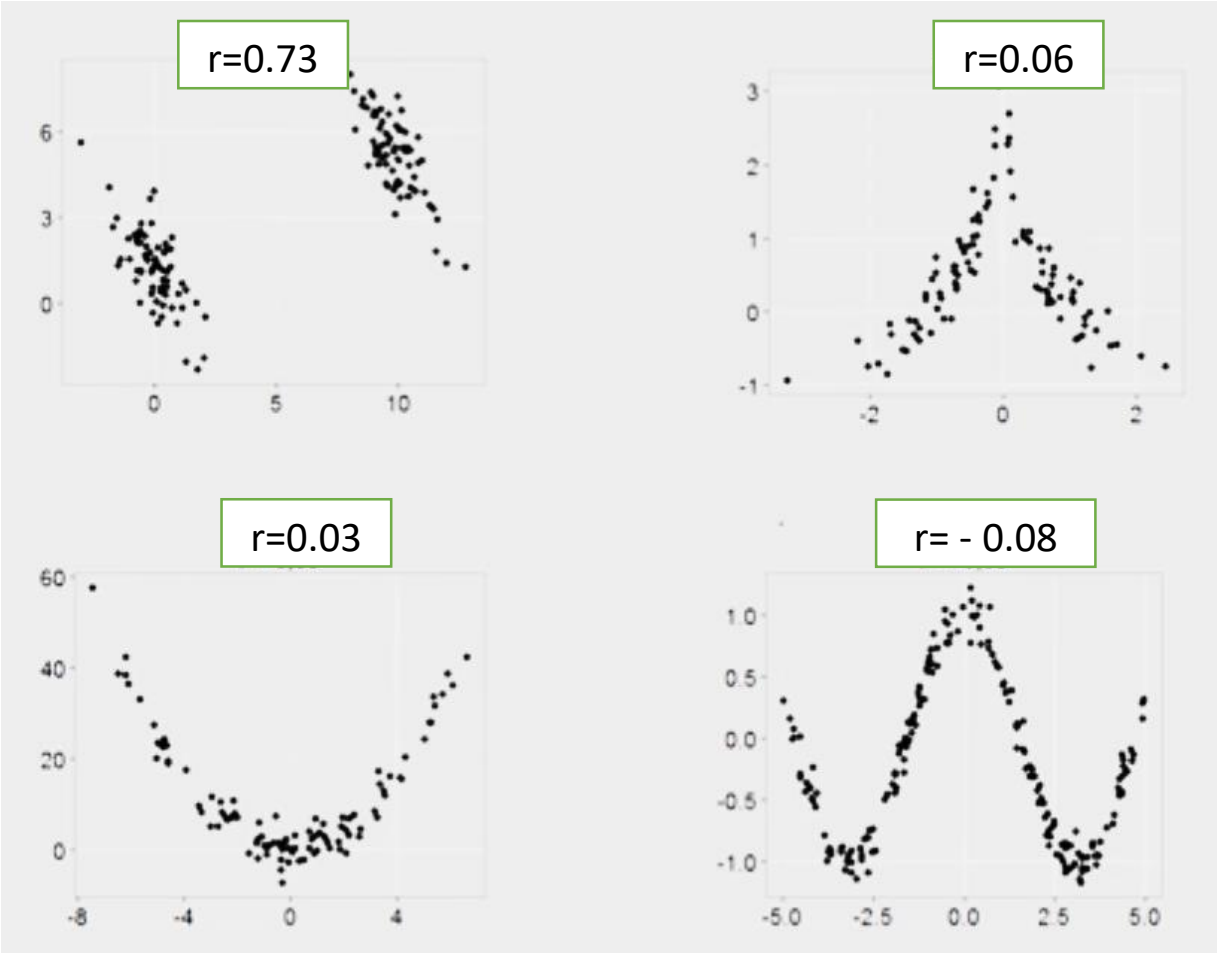
Словесная интерпретация величин коэффициента корреляции:

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

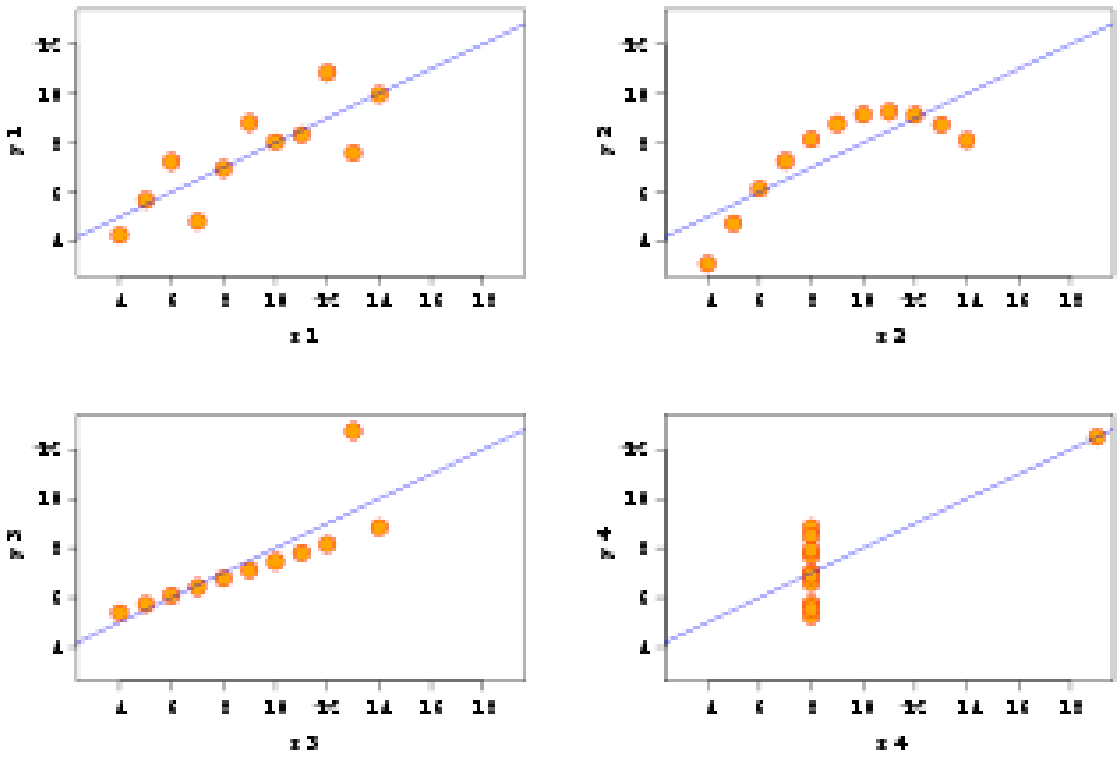
В качестве коэффициента корреляции между ранжированными переменными применяется **коэффициент Спирмена**, а для непрерывных переменных — **коэффициент корреляции Пирсона**.

У **коэффициента корреляции Пирсона** есть свои особенности и ограничения, которые лучше применять перед его использованием для анализа взаимосвязи двух переменных.

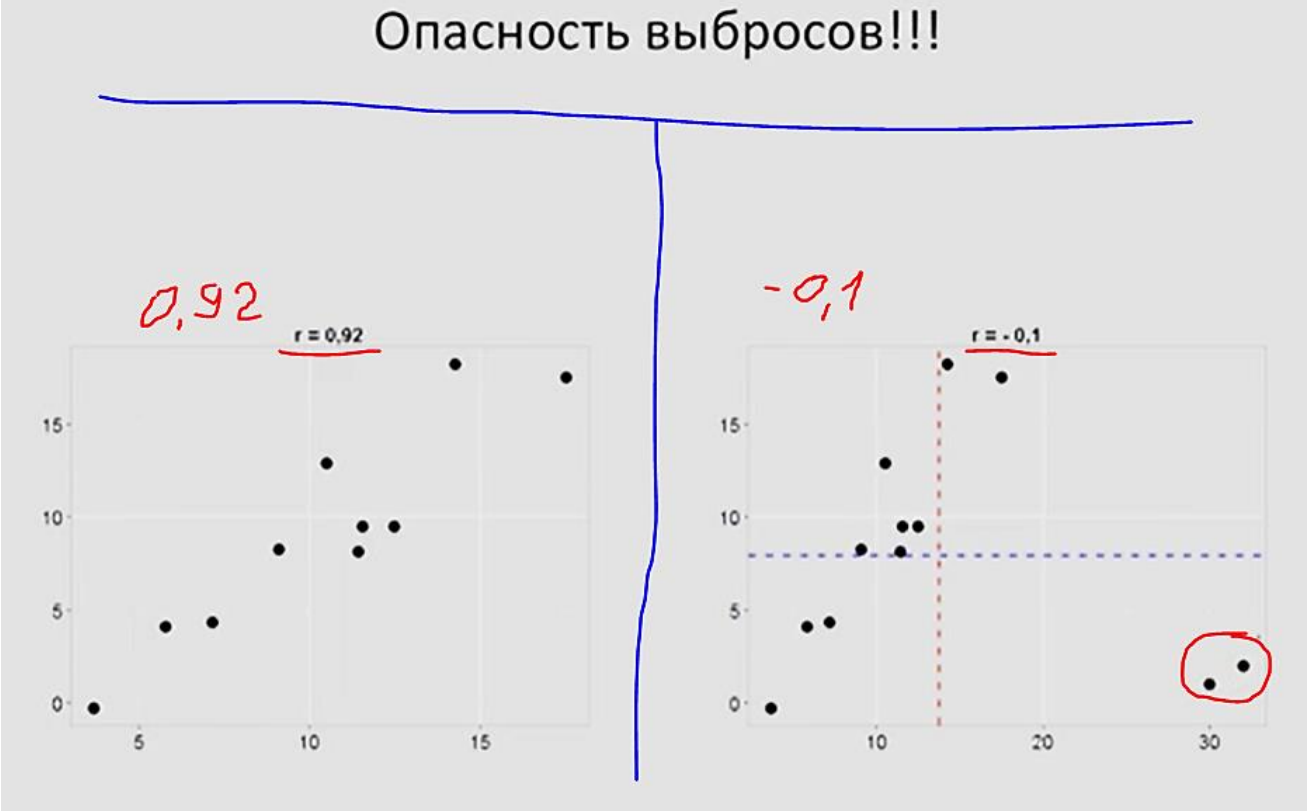
1. Характер связи должен быть линейным и монотонным, в противном случае применение коэффициента корреляции Пирсона будет некорректным:



Еще 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**

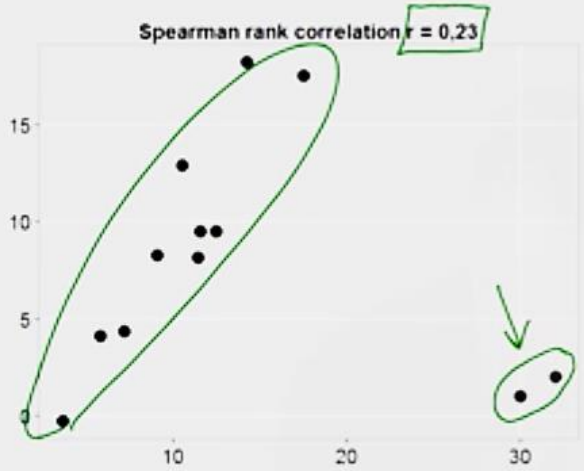
Коэффициент ранговой корреляции Спирмена [править | править код]

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.

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

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

Итог для выборки с выбросами
коэффициент корреляции Спирмена = 0.23
коэффициент корреляции Пирсона = -0.1

Коэффициент детерминации

R^2 — показывает, в какой степени дисперсия одной переменной обусловлена влиянием другой переменной

Равен квадрату коэффициента корреляции

Принимает значения [0, 1]

Вопрос: Что можно сказать о взаимосвязи этих двух величин?

Рост (сантиметры)	Заработная плата (рубли)	Коэффициент корреляции
167	32000	0,057338125
173	22000	
200	11000	
190	80000	
155	40000	
175	38000	
199	42000	
180	37000	
190	77000	
163	59000	

Линейная регрессия

Зависимость между двумя переменными может быть разной. Пример линейной зависимости показан на рис.1.

Рис. 1. Диаграмма разброса, исходные данные площади и годовые объемы продаж 14 магазинов 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 $x = (x_1, ..., x_r)$, где r – это число параметров объекта, предполагает, что между y и x линейное отношение:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon.$$

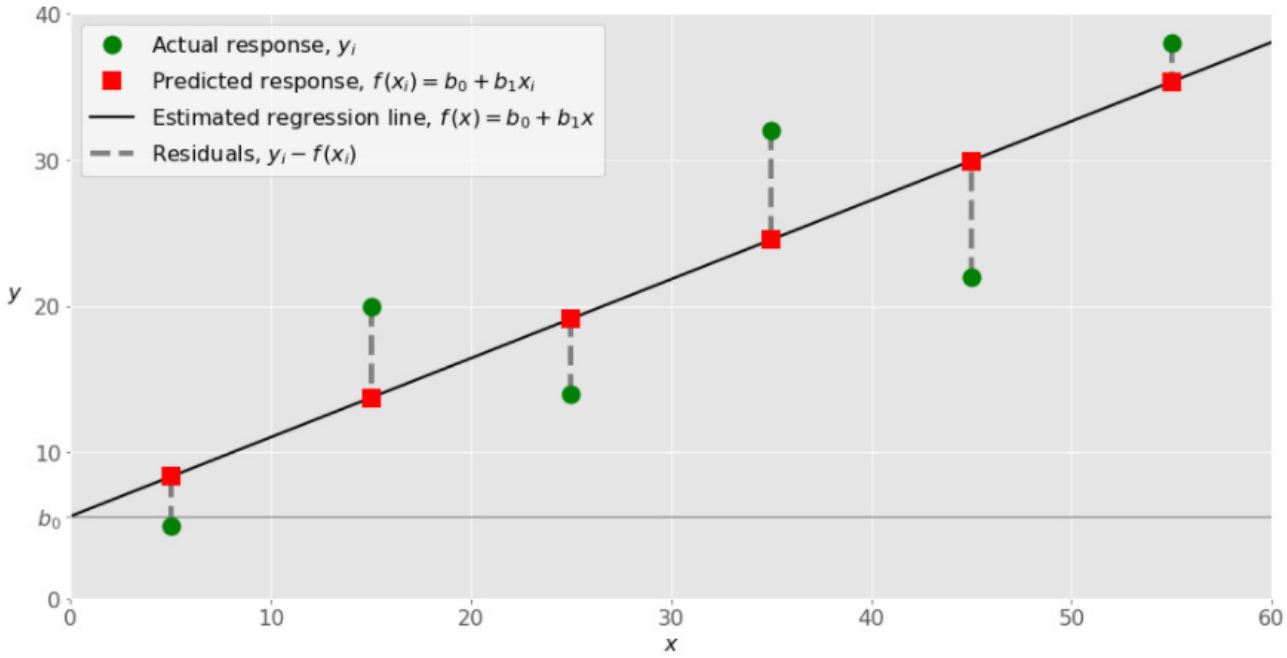
Это уравнение регрессии. β_0, β_1, …, β_r – коэффициенты регрессии, и ε – случайная ошибка.

Линейная регрессия вычисляет коэффициенты регрессии или просто прогнозируемые веса измерения, обозначаемые как b_0, b_1, …, b_r. Они определяют оценочную функцию регрессии

$$f(x) = b_0 + b_1 x_1 + \cdots + b_r x_r.$$

Простая линейная регрессия

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



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 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.

Для определения функции регрессии необходимо рассчитать оптимальные значения коэффициентов b_o 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**.

Суть метода наименьших квадратов (МНК).

Задача заключается в нахождении коэффициентов линейной зависимости, при которых функция двух переменных *a* и *b* $F(a,b)=\sum_{i=1}^{n}(y_i-(ax_i+b))^2$ принимает наименьшее значение. То есть, при данных *a* и *b* сумма квадратов отклонений экспериментальных данных от найденной прямой будет наименьшей. В этом вся суть метода наименьших квадратов.

Таким образом, решение примера сводится к нахождению экстремума функции двух переменных.

Вывод формул для нахождения коэффициентов

Составляется и решается система из двух уравнений с двумя неизвестными. Находим частные производные функции $F(a,b)=\sum_{i=1}^{n}(y_i-(ax_i+b))^2$ по переменным *a* и *b*, приравниваем эти производные к нулю.

$$\begin{cases}\frac{\partial F(a,b)}{\partial a}=0\\ \frac{\partial F(a,b)}{\partial b}=0\end{cases}\Leftrightarrow\begin{cases}-2\sum_{i=1}^{n}(y_i-(ax_i+b))x_i=0\\ -2\sum_{i=1}^{n}(y_i-(ax_i+b))=0\end{cases}\Leftrightarrow$$

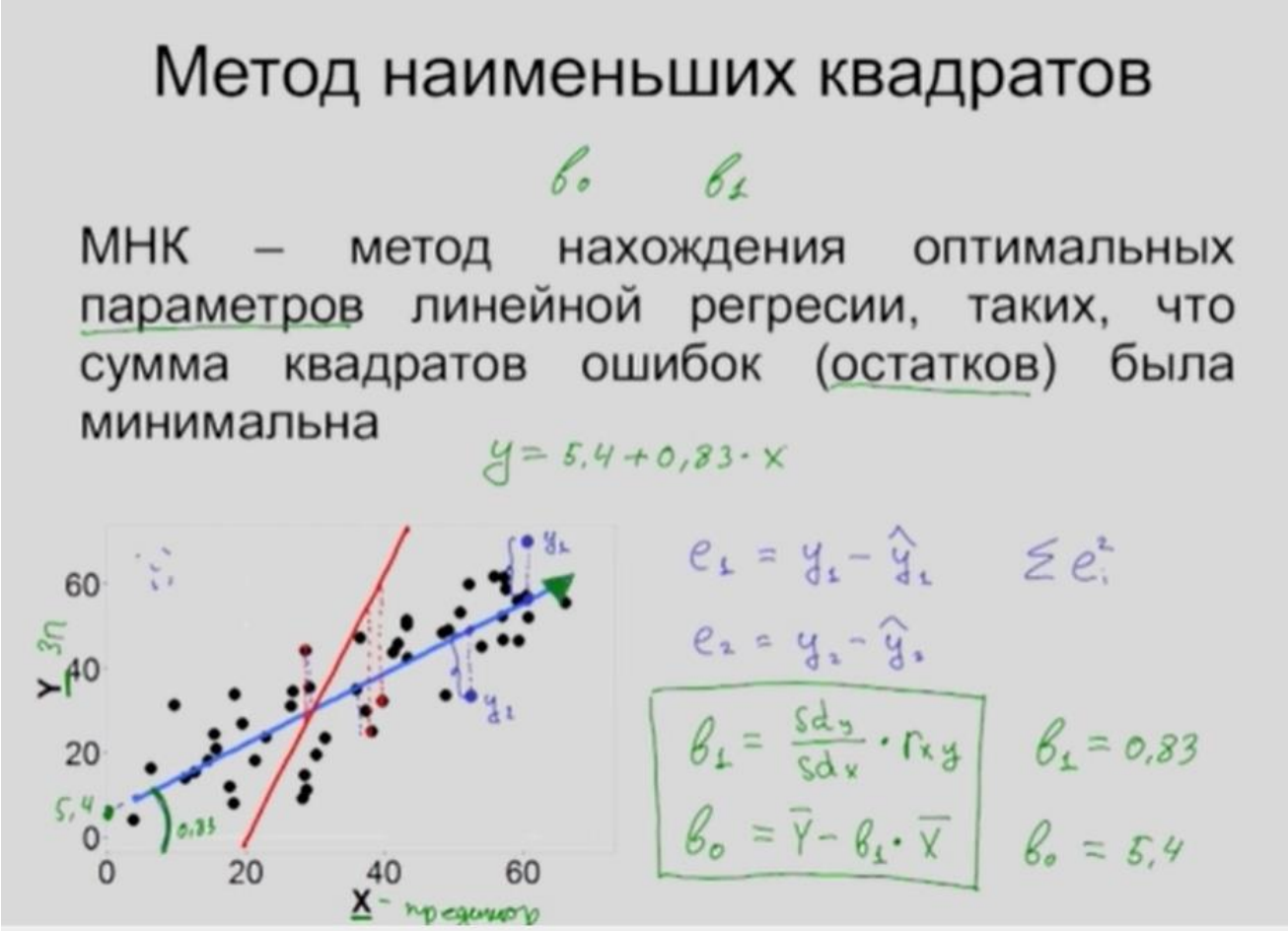
Почему?

$$\begin{cases}a\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_iy_i\\ a\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}b=\sum_{i=1}^{n}y_i\end{cases}\Leftrightarrow\begin{cases}a\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_iy_i\\ a\sum_{i=1}^{n}x_i+nb=\sum_{i=1}^{n}y_i\end{cases}$$

Решаем полученную систему уравнений любым методом (например *методом подстановки* или методом Крамера) и получаем формулы для нахождения коэффициентов по методу наименьших квадратов (МНК).

$$\begin{cases} a = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ b = \dfrac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n} \end{cases}$$

При данных *a* и *b* функция $F(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2$ принимает наименьшее значение.



Ограничения линейной регрессии

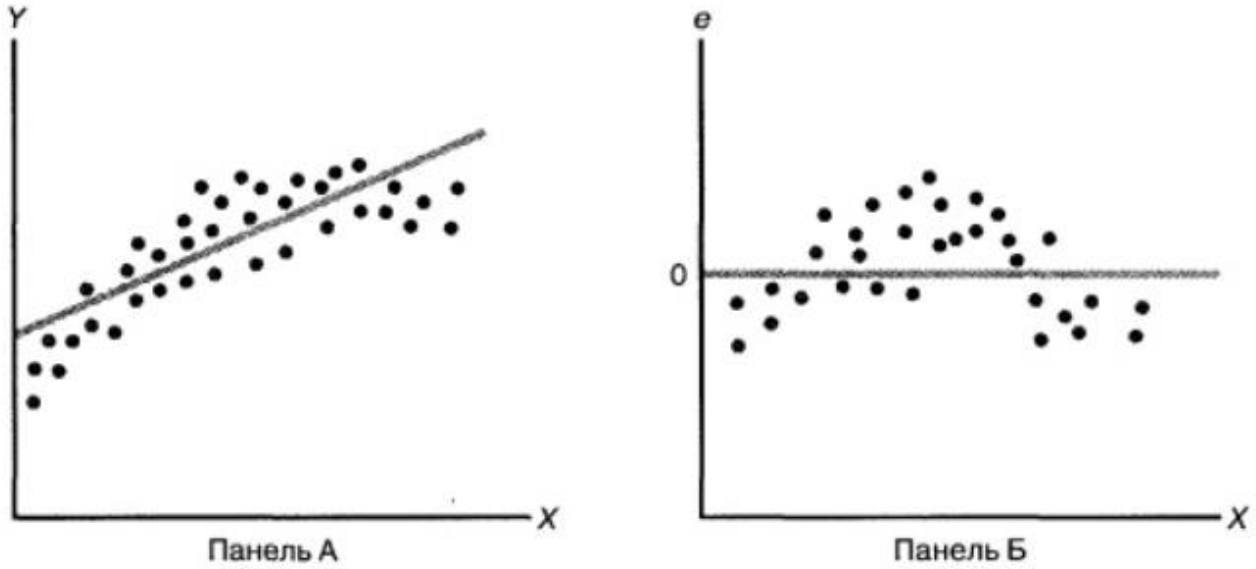
https://habr.com/ru/post/350668/

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

- ✓ Должно соблюдаться условие линейности

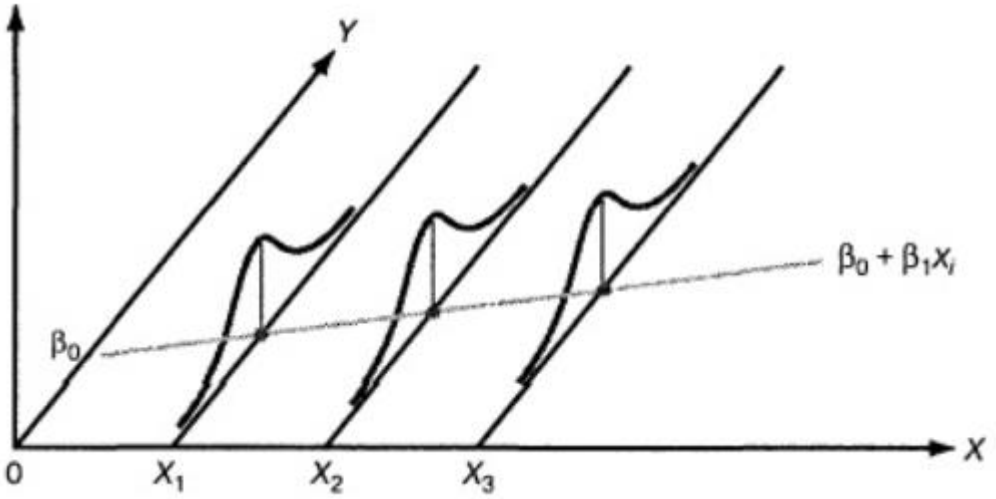
- ✓ Ошибка должна иметь нормальное распределение.
- ✓ Вариация данных вокруг линии регрессии должна быть постоянной.
- ✓ Ошибки должны быть независимыми.

Первое **предположение о линейности** предполагает корректный выбор модели регрессии. При соблюдении условия линейности увеличение, или уменьшение вектора независимых переменных в k раз, приводит к изменению зависимой переменной также в k раз.



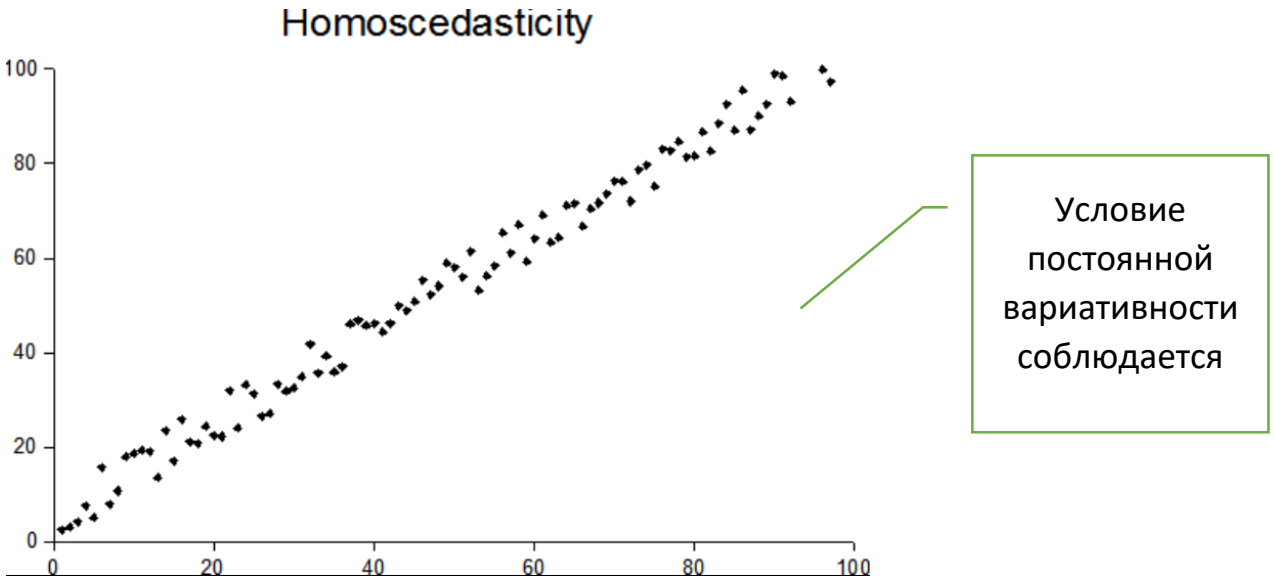
Пример: Панель А иллюстрирует возрастание переменной *Y* при увеличении переменной *X*. Однако зависимость между этими переменными носит нелинейный характер, поскольку скорость возрастания переменной *Y* падает при увеличении переменной *X*. Таким образом, для аппроксимации зависимости между этими переменными лучше подойдет квадратичная модель. Особенно ярко квадратичная зависимость между величинами X_i и e_i проявляется на панели Б. Графическое изображение позволяет отфильтровать или удалить линейную зависимость между переменными *X* и *Y* и выявить недостаточную точность модели простой линейной регрессии. Таким образом, в данной ситуации вместо простой линейной модели должна применяться квадратичная модель, обладающая более высокой точностью.

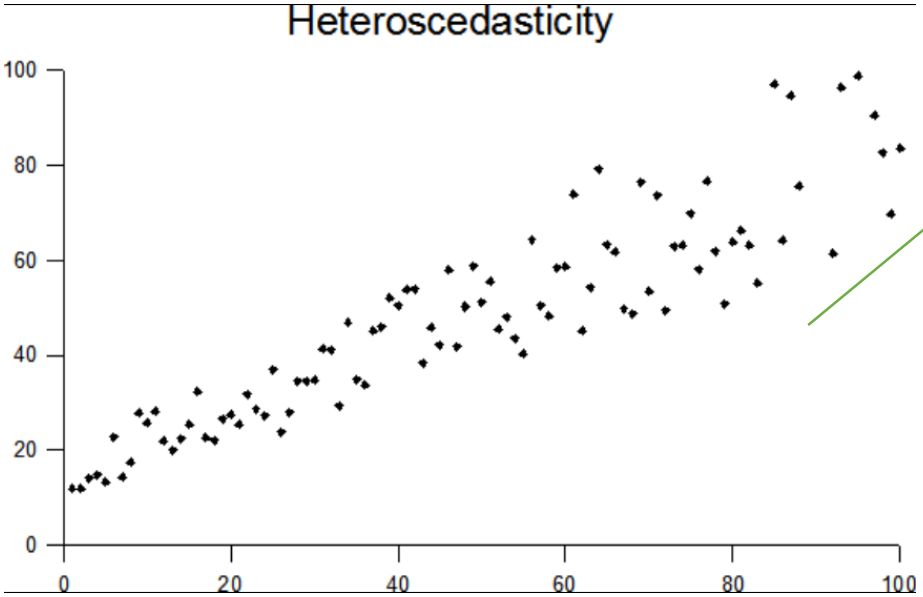
Второе **предположение, о нормальном распределении ошибок**, требует, чтобы при каждом значении переменной X ошибки линейной регрессии имели нормальное распределение (рис.). Регрессионный анализ довольно устойчив к незначительным нарушениям этого условия. Если распределение ошибок относительно линии регрессии при каждом значении X не слишком сильно отличается от нормального, выводы относительно линии регрессии и коэффициентов регрессии изменяются незначительно.



Предположение о нормальном распределении ошибок

Второе условие заключается в том, что **вариация данных вокруг линии регрессии должна быть постоянной при любом значении переменной X**. Это означает, что величина ошибки как при малых, так и при больших значениях переменной X должна изменяться в одном и том же интервале (см. рис. 7). Это свойство очень важно для метода наименьших квадратов, с помощью которого определяются коэффициенты регрессии. Если это условие нарушается, следует применять либо преобразование данных, либо метод наименьших квадратов с весами.



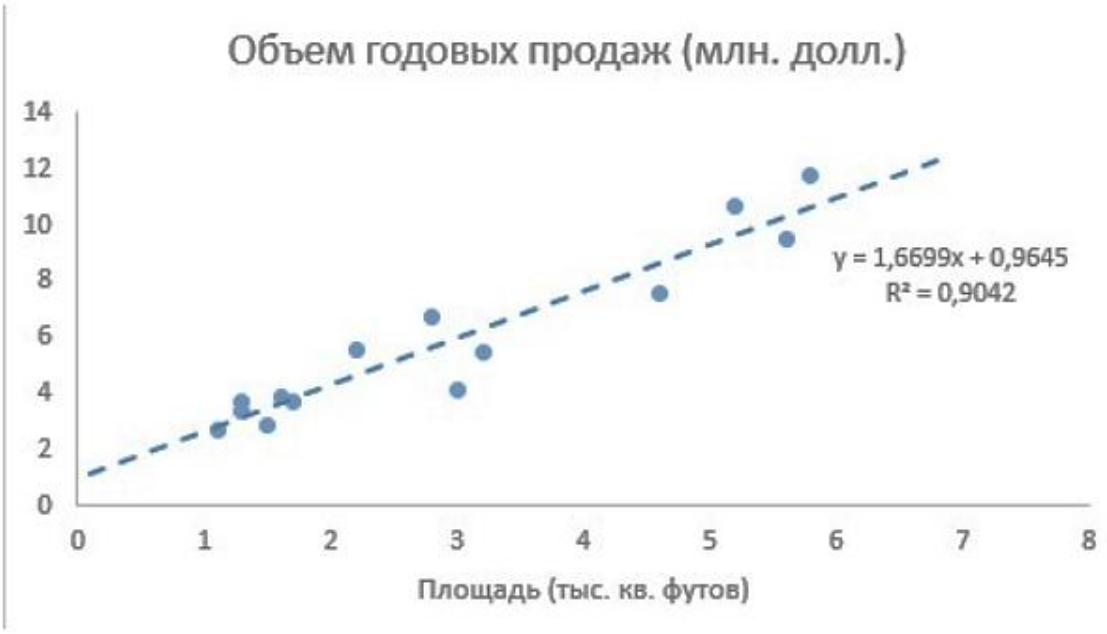


Условие постоянной вариативности нарушено

Третье предположение, о независимости ошибок, заключается в том, что ошибки регрессии не должны зависеть от значения переменной *X* и не должны коррелировать с другими ошибками. Это условие особенно важно, если данные собираются на протяжении определенного отрезка времени. В этих ситуациях ошибки, присущие конкретному отрезку времени, часто коррелируют с ошибками, характерными для предыдущего периода.

Прогнозирование в регрессионном анализе: интерполяция и экстраполяция

Применяя регрессионную модель для прогнозирования, необходимо учитывать лишь допустимые значения независимой переменной. В этот диапазон входят все значения переменной X, начиная с минимальной и заканчивая максимальной. Таким образом, предсказывая значение переменной Y при конкретном значении переменной X, исследователь выполняет интерполяцию между значениями переменной X в диапазоне возможных значений. Однако **экстраполяция значений за пределы этого интервала не всегда релевантна**. Например, пытаясь предсказать среднегодовой объем продаж в магазине, зная его площадь (рис. 3а), мы можем вычислять значение переменной Y лишь для значений X от 1,1 до 5,8 тыс. кв. футов.



Следовательно, прогнозировать среднегодовой объем продаж можно лишь для магазинов, площадь которых не выходит за пределы указанного диапазона. Любая попытка экстраполяции означает, что мы предполагаем, будто линейная регрессия сохраняет свой характер за пределами допустимого диапазона.

Множественная линейная регрессия

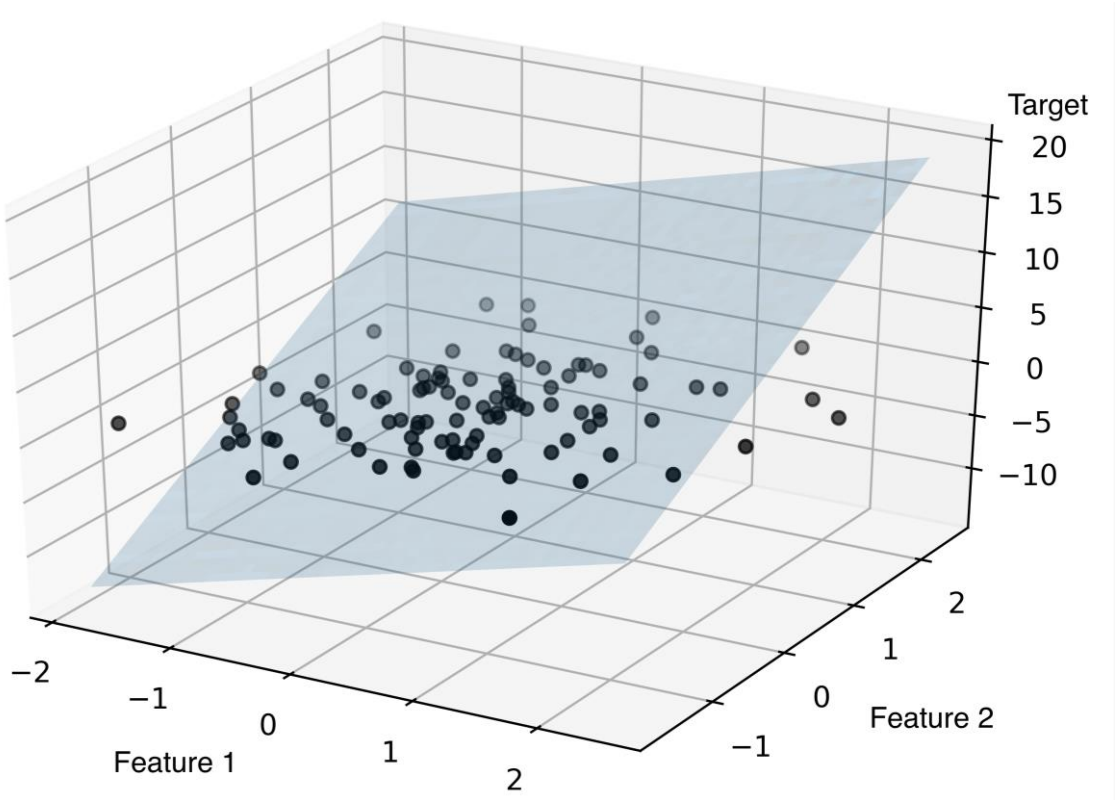
Множественной называют линейную регрессию, в модели которой число независимых переменных две или более.

Уравнение множественной линейной регрессии имеет вид:

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

Как и в простой линейной регрессии, параметры модели b_n вычисляются при помощи метода наименьших квадратов.

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



Преимущество множественной линейной регрессии по сравнению с простой заключается в том, что **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**.

Однако в множественной линейной регрессии возникают и проблемы, нехарактерные для простой модели:

- ✓ возможно появление мультиколлениарности;
- ✓ необходимо выбирать лучшую модель, в которой минимальный набор независимых переменных сможет объяснить наибольшую долю дисперсии зависимой. Для этих целей используется различные информационные критерии.