Эконометрика

Вторая лекция

Модель парной регрессии

Простейшая регрессионная модель:

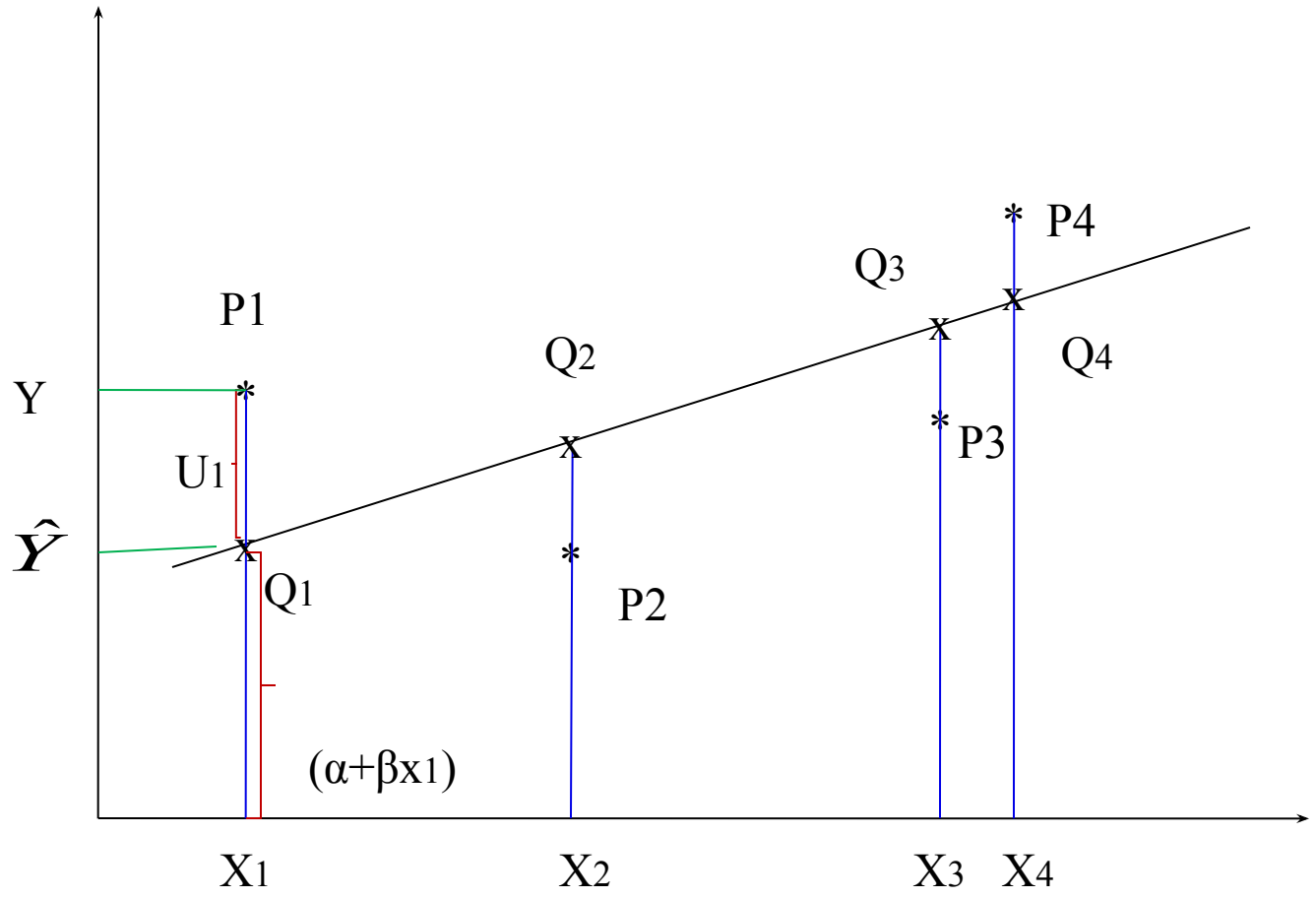
$$y=\alpha+\beta x+u$$

y - зависимая переменная, объясняемая, регрессант

x – независимая переменная, объясняющая, регрессор

α и **β** — параметры модели

U – случайная составляющая


P4
Q3
P1
Q2
Q4
Y
P3
U1
$\hat{Y}$
Q1
P2
(α+βx1)
X1
X2
X3
X4

Величина y - зависимая переменная, состоит из двух частей:

1)Неслучайной составляющей – **$(\alpha+\beta x)$,**

2) Случайной составляющей - **u.**

Точки **P_1, P_2, P_3** и **P_4** – это фактические или наблюденные значения.

Точки **Q_1, Q_2, Q_3** и **Q_4** – это теоретические значения, т.е. в отсутствии случайной компоненты.

Задача регрессионного анализа состоит в нахождении оценок **α** и **β** и в определении положения регрессионной прямой по известным или наблюденным значениям **X** и **Y** при неизвестных значениях **U**

Метод наименьших квадратов

- **МНК** является наиболее популярным методом нахождения оценок неизвестных параметров. Критерий выбора наилучших параметров: минимизация суммы квадратов остатков.

- **Остаток или отклонение (e)** – разница между наблюдаемым значением переменной **Y** и ее теоретическим значением $\hat{Y}$ в каждом наблюдении, т.е. при каждом значении **X**.

$$1) \quad x_1 \rightarrow (Y_1 - \hat{Y}_1)^2 = e_1^2$$

$$2) \quad x_2 \rightarrow (Y_2 - \hat{Y}_2)^2 = e_2^2$$

$$3) \quad x_3 \rightarrow (Y_3 - \hat{Y}_3)^2 = e_3^2$$

$$\vdots$$

$$n) \quad x_n \rightarrow (Y_n - \hat{Y}_n)^2 = e_n^2$$

Критерий оптимизации:

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 =$$

Предположим, что между Y и X существует прямая связь, т.е.

$$\hat{Y}_i = a + bx_i$$

Тогда можно записать:

$$= \sum_{i=1}^{n} (Y_i - a - bx_i)^2 \rightarrow \min$$

Функция принимает свое минимальное значение в точке, где все ее частные производные равны нулю

$$\begin{cases} \dfrac{\partial S}{\partial a} = \boxtimes = 0 \\ \dfrac{\partial S}{\partial b} = \boxtimes = 0 \end{cases}$$

Данная система называется системой нормальных уравненний, решая эту систему относительно **a** и **b**, мы получаем рабочие формулы для нахождения оценок неизвестных параметров **α** и **β** исходного уравнения.

Формулы для нахождения оценок a и b:

$$a = \overline{y} - b \cdot \overline{x}$$

$$b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}$$

Причины существования случайной компоненты

1. Не включение объясняющих переменных

Соотношение между **y** и **x** - очень большое упрощение. Существуют и другие факторы, влияющие на **y**, которые не учтены в модели. Влияние этих факторов приводит к тому, что наблюдаемые точки лежат вне прямой.

- Невозможность измерения.
- Слабое влияние фактора.
- Отсутствия опыта или знаний.

2. Агрегирование переменных

Во многих случаях зависимость — это попытка объединить вместе некоторое число микроэкономических соотношений. Отдельные соотношения имеют разные параметры, любая попытка определить соотношение между ними является лишь аппроксимацией.

3. Неправильное описание структуры модели

Если зависимость относится к данным о временном ряде, то значение **Y** может зависеть не от фактического значения **X**, а от значения, которое ожидалось в предыдущем периоде.

Если ожидаемое и фактическое значения тесно связаны, то будет казаться, что между **Y** и **X** существует зависимость, но это будет лишь аппроксимация.

4. Неправильная функциональная спецификация

Функциональное соотношение между **Y** и **X** математически может быть определено неправильно. Истинная зависимость может не являться линейной, а быть более сложной.

5. Ошибки измерения

Если в измерении одной или более взаимосвязанных переменных имеются ошибки, то наблюдаемые значения не будут соответствовать точному соотношению.

Случайная компонента является суммарным проявлением всех факторов. Если бы случайной компоненты не существовало, то мы бы знали, что любое изменение **Y** вызвано только изменением **X** и смогли бы точно вычислить β.

Однако в действительности каждое изменение Y отчасти вызвано изменением U. Поэтому мы не можем вычислить истинные значения параметров (**α** и **β**), а можем определить лишь их оценки, т.е. приближенные значения (**a** и **b**).

Свойства коэффициентов регрессии и условия нормальной линейной регрессии (Гаусса-Маркова)

Фактическое значение **Y** состоит из двух элементов: из *неслучайной части* и *случайной компоненты*, поэтому вычисленные оценки ***a*** и ***b*** также состоят из двух элементов. Неслучайной частью для ***a*** является **α**, для ***b*** **– β.**

Следовательно, свойства коэффициентов регрессии существенным образом зависят от свойств *случайной компоненты*.

Для того чтобы регрессионный анализ, основанный на обычном МНК, давал наилучшие результаты, случайный член должен удовлетворять четырем условиям, известным как ***условия Гаусса—Маркова.***

1-е условие Гаусса—Маркова

Математическое ожидание случайной компоненты в любом наблюдении должно быть равно нулю. Иногда величина случайной компоненты будет положительной, иногда отрицательной, но она не должена иметь систематического смещения ни в одном из двух возможных направлений.

Фактически если уравнение регрессии включает константу, то можно предположить, что это условие выполняется автоматически, так как роль константы состоит в определении любой систематической тенденции в поведении **Y**, которую не учитывают объясняющие переменные, включенные в уравнение регрессии.

2-е условие Гаусса—Маркова

Дисперсия σ_u^2 случайной компоненты должна быть постоянна для всех наблюдениях. Иногда случайная компонента будет больше, иногда меньше, однако не должно быть априорной причины для того, чтобы она порождала большую ошибку в одних наблюдениях, чем в других.

Если это условие выполняется, то говорят, что дисперсия ошибки ***гомоскедастична***,

если нет, то - ***гетероскедастична***.

3-е условие Гаусса—Маркова

Даное условие предполагает отсутствие систематической связи между значениями случайной компоненты в любых двух наблюдениях. Например, если случайная компонента велика и положительна в одном наблюдении, это не должно обусловливать систематическую тенденцию к тому, что она будет большой и положительной в следующем наблюдении

Или большой и отрицательной, или малой и положительной, или малой и отрицательной.

Случайные компоненты должна быть абсолютно независимы друг от друга.

Выполнение данного условия гарантирует отсутствие автокорреляции.

В противном случае, говорят, что случайная компонета автокоррелирована.

4-е условие Гаусса—Маркова:

- Случайная компонента должна быть распределена независимо от объясняющих переменных.

Предположение о нормальности:

Наряду с условиями Гаусса—Маркова обычно также предполагается нормальность распределения случайного члена. Если случайный член нормально распределен, то так же будут распределены и коэффициенты регрессии. Предположение о нормальности основывается на центральной предельной теореме:

«Если случайная величина является общим результатом взаимодействия большого числа других случайных величин, ни одна из которых не является доминирующей, то она будет иметь приблизительно нормальное распределение, даже если отдельные составляющие не имеют нормального распределения».

Интерпретация линейного уравнения регрессии

$$\hat{y}=a+bx$$

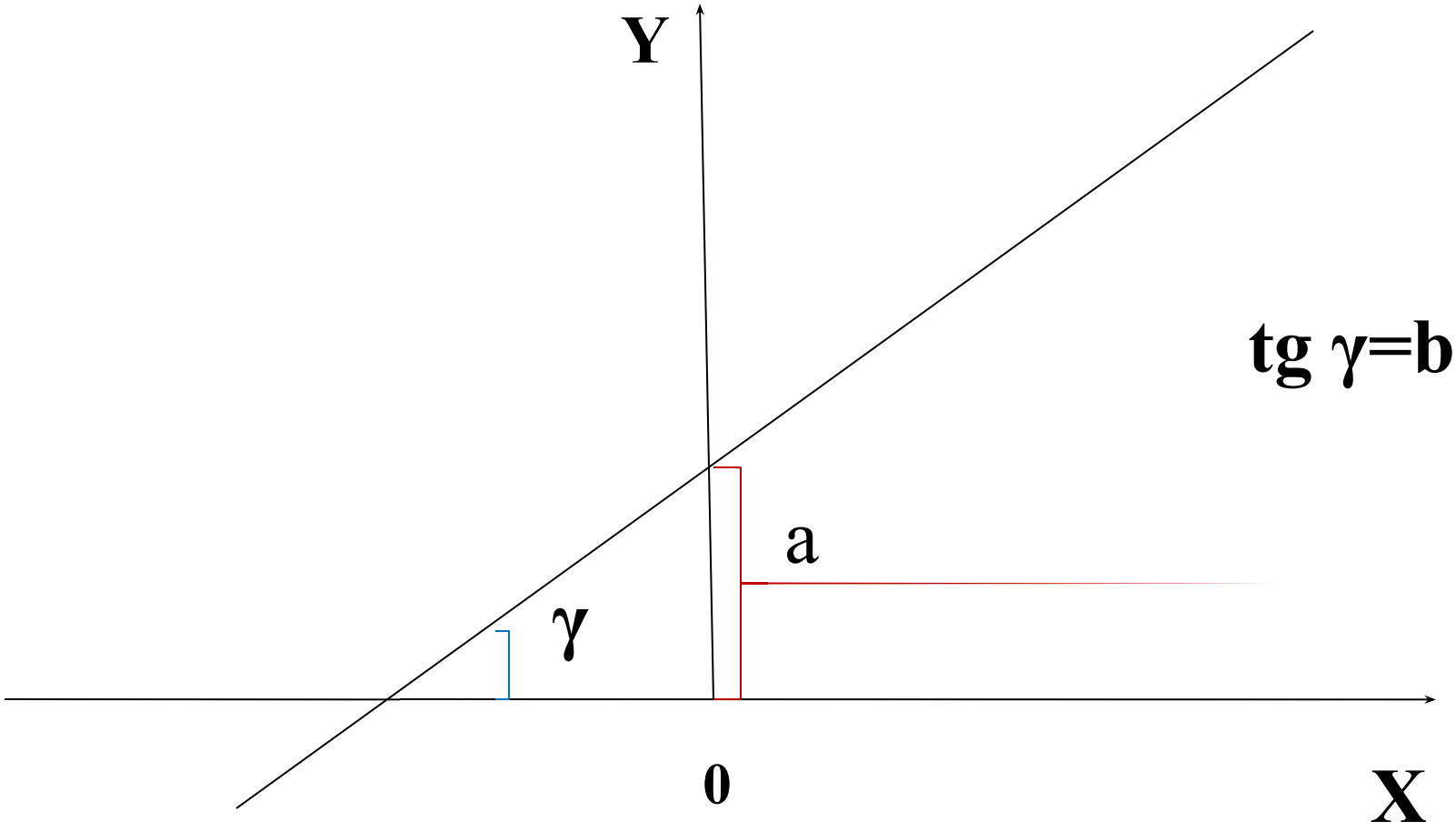
Оценки a и b имеют математическую и экономическую интерпретацию.

Математическая:

Коэффициент a называется регрессионной постоянной или **const.** Это значение $\hat{y}$, в том случае когда **x=0**.

Геометрически это точка с координатами: **(0,a)**

Коэффициент **b** – коэффициент регрессии – это тангенс угла наклона к оси ОХ.


Y
tg γ=b
a
γ
0
X

Экономическая:

a – регрессионная постоянная, const

Дает прогнозное значение **y**, в том случае, когда факторный признак равен нулю. Экономически это может иметь или не иметь ясного смысла.

b – коэффициент регрессии

Показывает на сколько изменится значение **y** (в единицах измерения **y**), если **x** возрастет на одну единицу (в единицах измерения **x**) от своего среднего уровня.

По группе предприятий, выпускающих один и тот же вид продукции, рассматривается функция издержек

$$y=\alpha+\beta x+u ,$$

где **X** – объем выпуска продукции (тыс.шт.),

Y – затраты на производство (млн.руб.)

Номер предприятия	Выпуск продукции, тыс.ед., x	Затраты на производство млн. руб., Y
1	1	30
2	2	70
3	4	150
4	3	100
5	5	170
6	3	100
7	4	150
Итого	22	770

Оценив параметры модели методом наименьших квадратов, получим следующее уравнение:

$$\hat{Y} = -5{,}79 + 36{,}84 \cdot x$$

В данном случае величина параметра a не имеет экономического смысла.

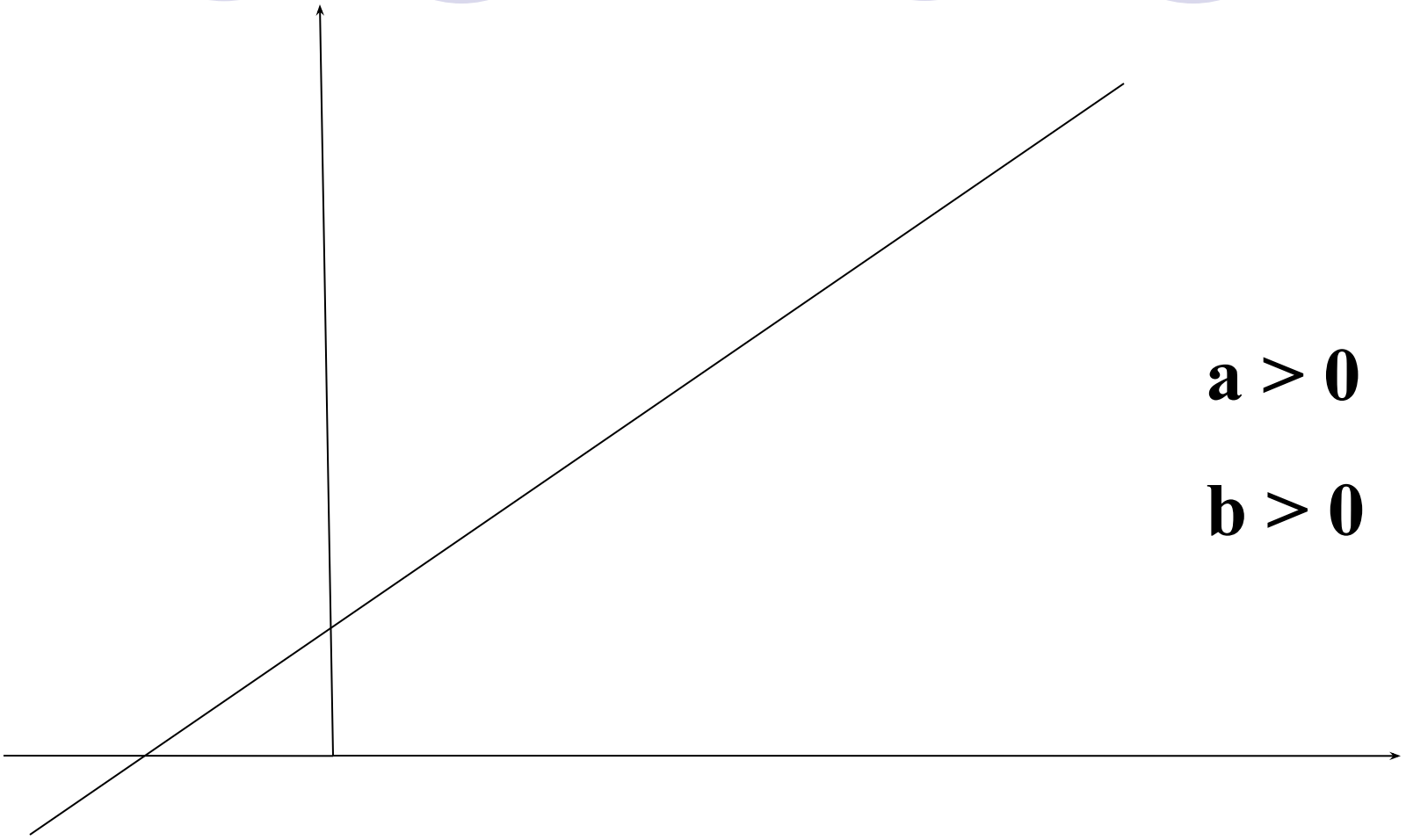
Параметр **b** показывает, что если выпуск продукции возрастет на одну тысячу штук (от своего среднего уровня), то затраты на производство увеличатся на 36,84 млн.руб.

Суточное потребление сливочного масла в обследованных семьях связано с доходом потребителя прямолинейной регрессией следующим образом:

$$\hat{y}=3,87+0,418*x$$

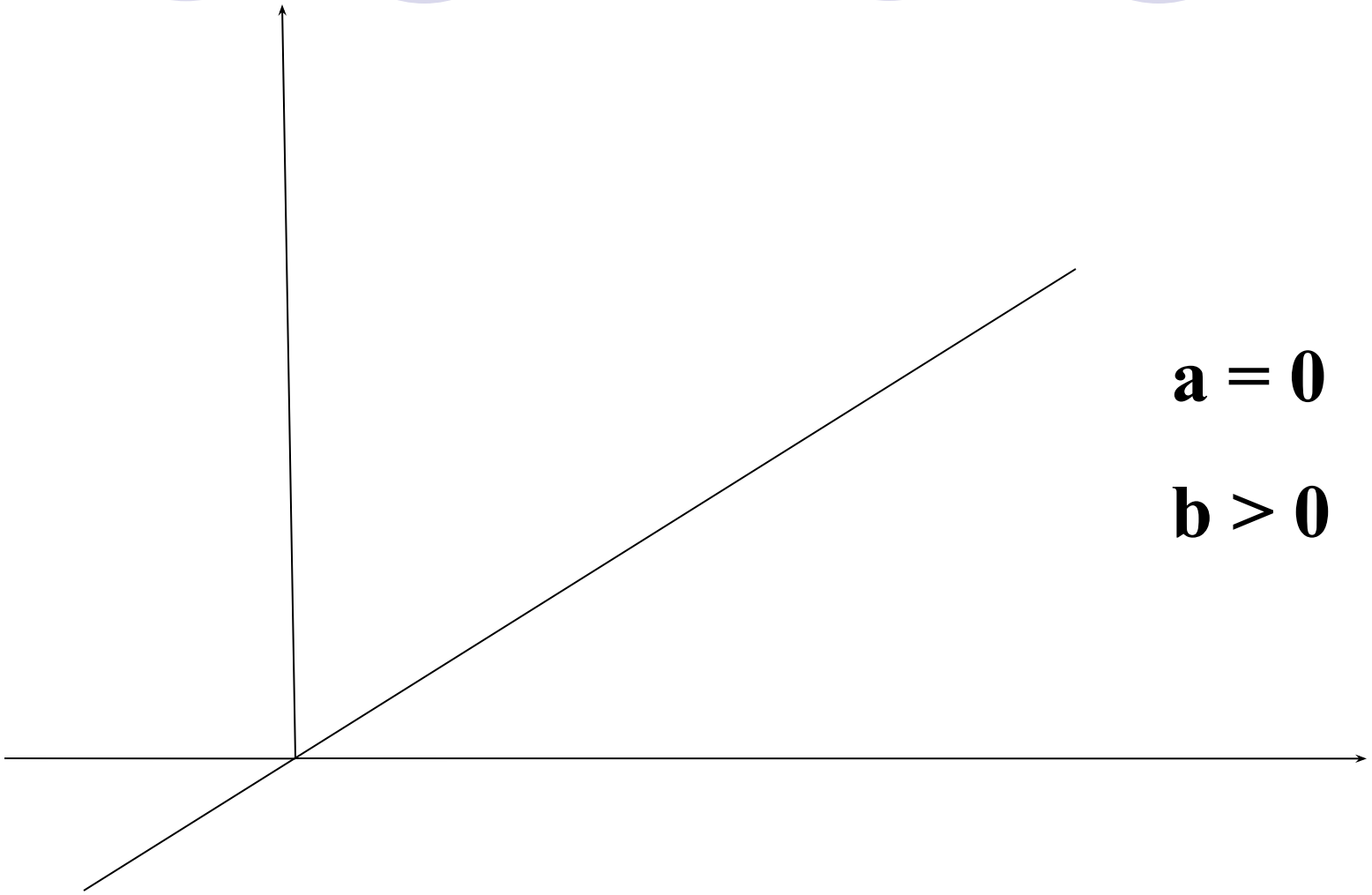
x – доход (руб.)

y – сливочное масло (г/сут.)



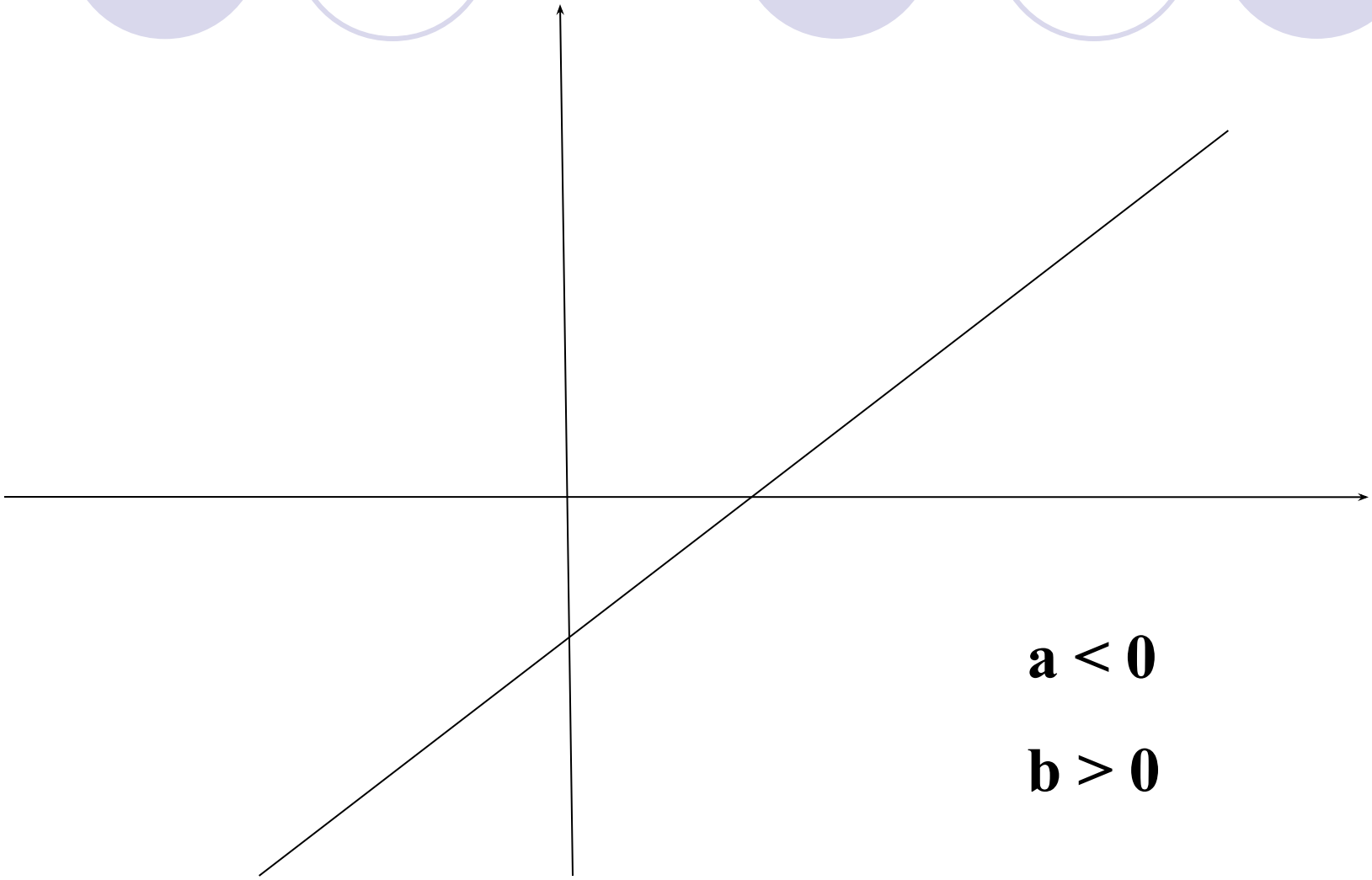
a > 0

b > 0



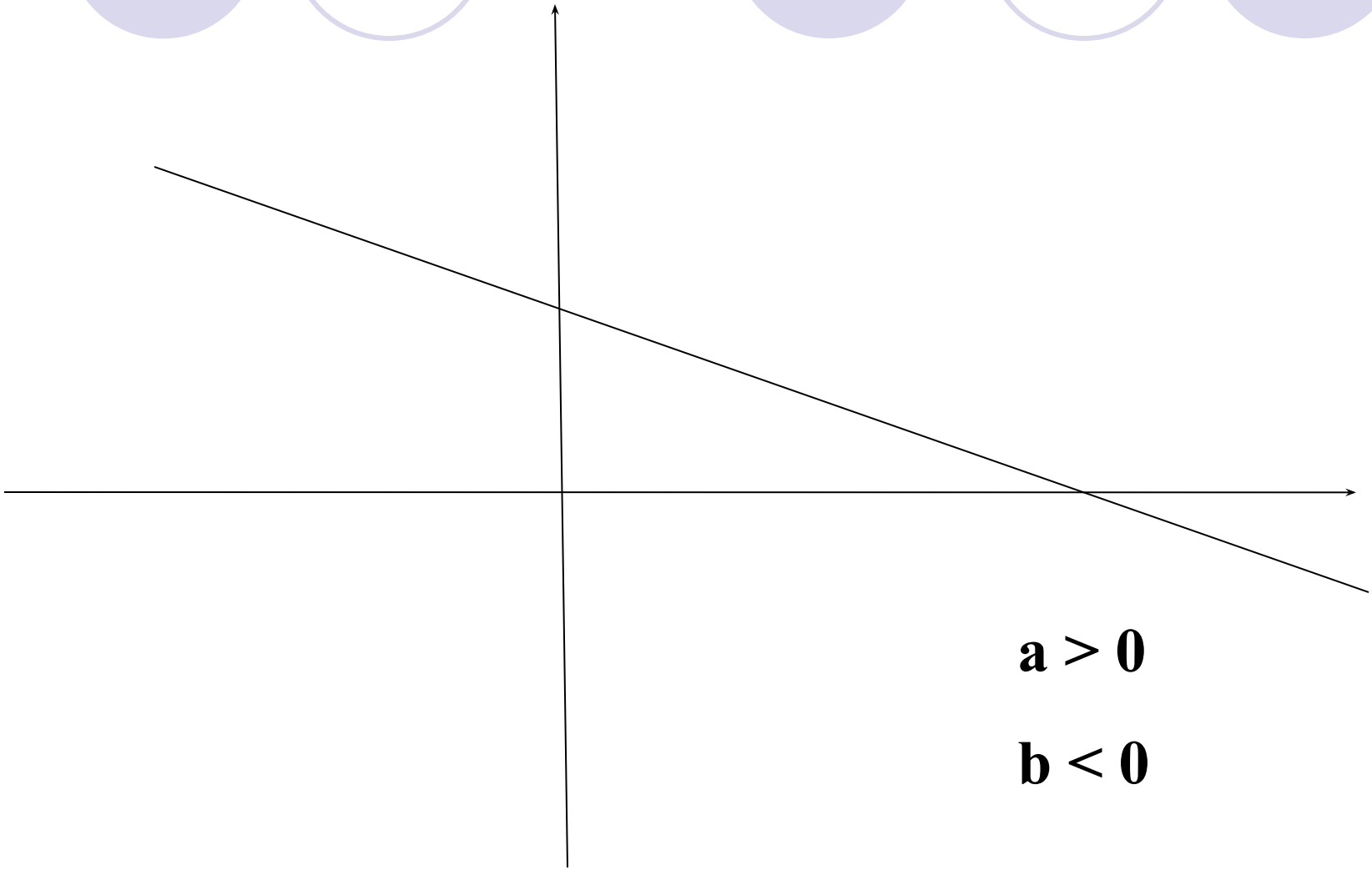
$a = 0$

$b > 0$



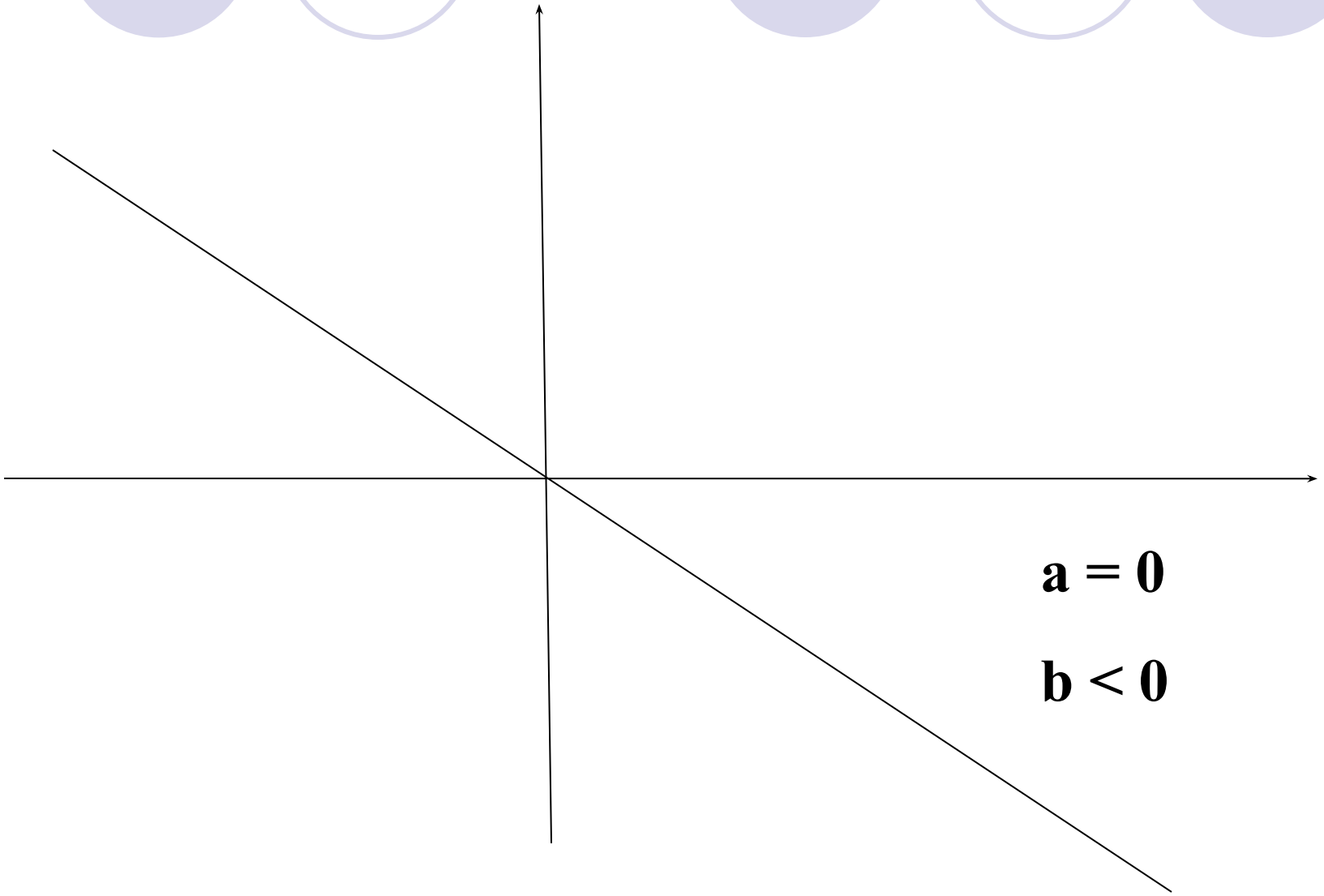
a < 0

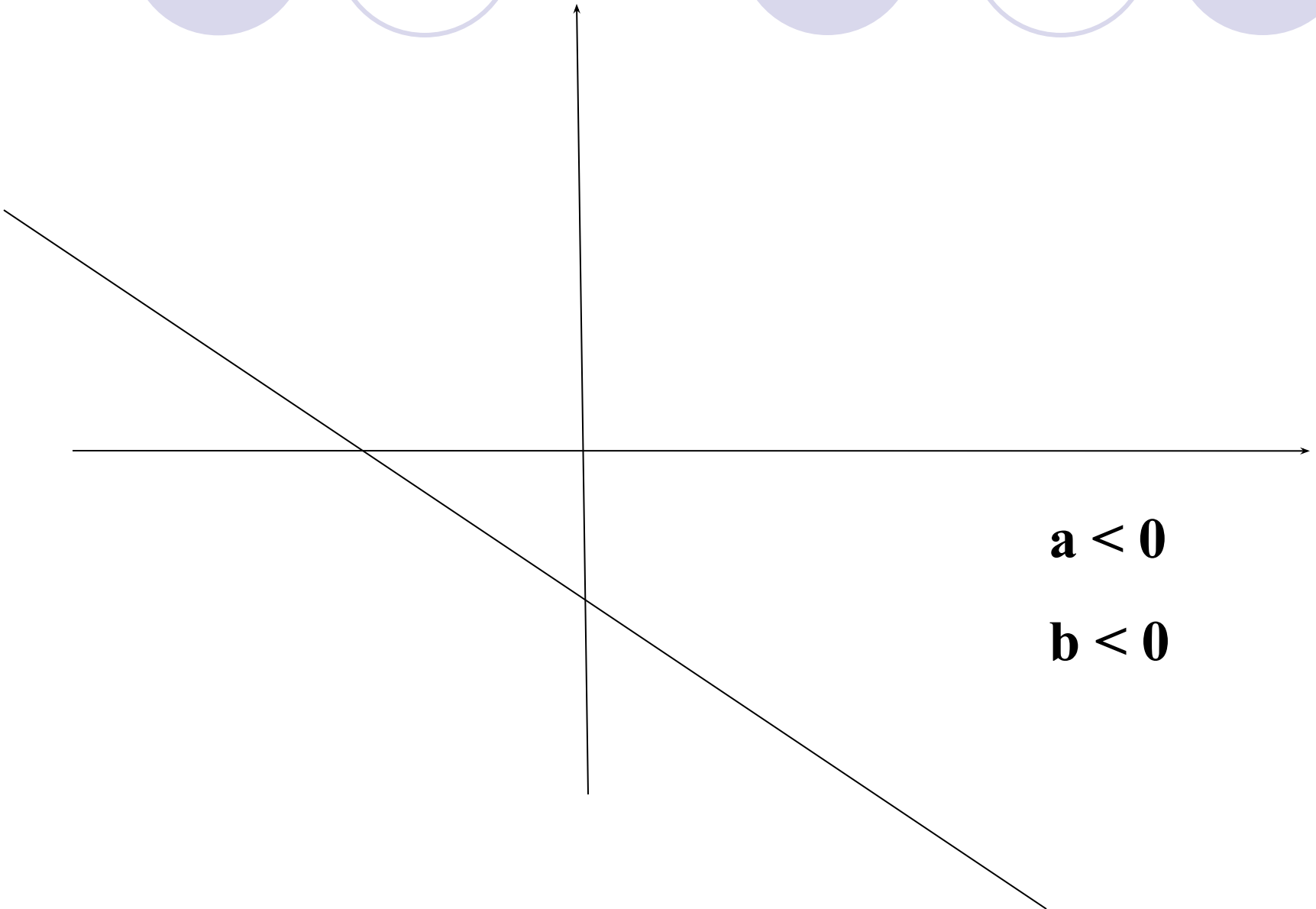
b > 0



a > 0

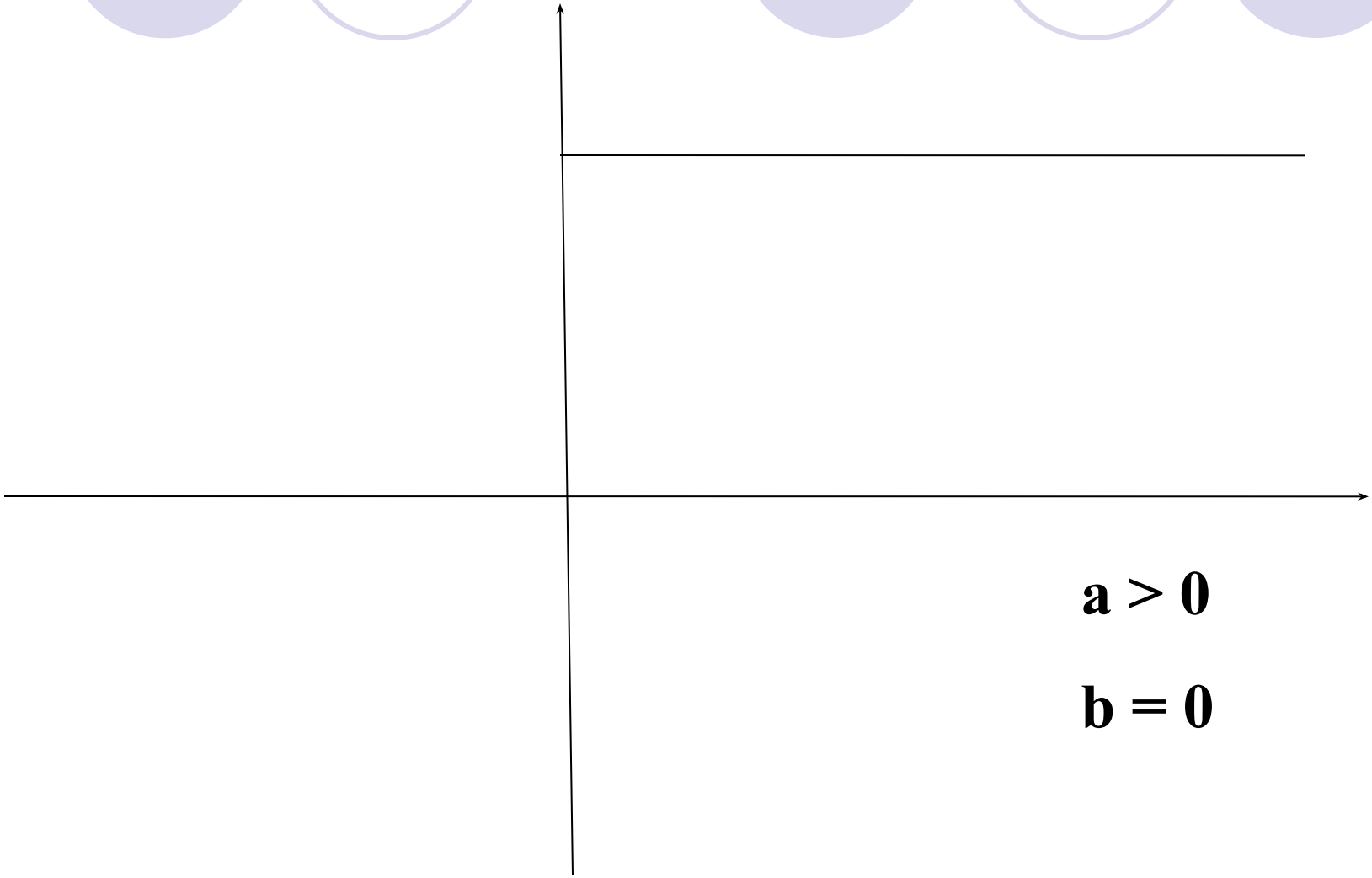
b < 0


a = 0
b < 0



a < 0

b < 0



$a > 0$

$b = 0$

Определение тесноты связи между факторами

В качестве меры тесноты связи используется линейный коэффициент корреляции:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

где

$$\sigma_x = \sqrt{\overline{x^2} - (\overline{x})^2}$$

$$\sigma_y = \sqrt{\overline{y^2} - (\overline{y})^2}$$

Линейный коэффициент корреляции может принимать любые значения в пределах от минус 1 до плюс 1. Чем ближе коэффициент корреляции по абсолютной величине к 1, тем теснее связь между признаками.

Знак при линейном коэффициенте корреляции указывает на направление связи - прямой зависимости соответствует знак плюс, а обратной зависимости - знак минус.

Если сравнить формулы для расчета коэффициентов регрессии и корреляции, то можно увидеть, что между этими коэффициентами существует связь

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y}$$

Можно выразить коэффициент корреляции через коэффициент регрессии:

$$r = b \cdot \frac{\sigma_x}{\sigma_y}$$

Если $b < 0$ $\Rightarrow -1 \leq r < 0$

Если $b > 0$ $\Rightarrow 0 < r \leq 1$

$r = 0$ ==> связь между x и y отсутствует

$0 < |r| \leq 0{,}3$ => связь практически отсутствует

$0{,}3 < |r| \leq 0{,}5$ => слабая связь между x и y.

$0{,}5 < |r| \leq 0{,}7$ => средняя (умеренная связь).

$0{,}7 < |r| < 1$ => сильная связь.

$|r| = 1$ => функциональная связь.

d – коэффициент детерминации.

Коэффициент детерминации показывает на сколько процентов изменение **y** обусловлено изменением **x**.

$$d = r^2 * 100\%$$

Оставшаяся доля приходится на влияние прочих факторов, не учтенных в модели.

$$(1 - r^2) * 100$$

Для интерпретации полученных результатов можно также использовать коэффициент эластичности, который показывает насколько процентов в среднем изменится значение результативного признака, если факторный признак увеличится на один процент.

$$Э = f'(x)\frac{x}{f(x)}$$

$$Э = b\frac{x}{a + b \cdot x}$$

В силу того, что коэффициент эластичности для линейной функции не является величиной постоянной, а зависит от соответствующего значения **X**, то обычно рассчитывается средний показатель эластичности по формуле:

$$Э = b \frac{\bar{x}}{a + b \cdot \bar{x}}$$

В нашем примере коэффициент эластичности равен 1,03 %.

Это означает, что с ростом выпуска продукции на 1 % затраты на производство в среднем увеличатся на 1,03 %.

- Коэффициент корреляции также как и коэффициент регрессии должен быть подвергнут оценке статистической значимости. Для этого, сначала рассмотрим разложение общей дисперсии на объяснимую (факторную) и необъяснимую (остаточную).
- коэффициент корреляции статистически значим, если:

$$\sigma^2_{общ} = \sigma^2_{факт} + \sigma^2_{ост}$$

$$r^2 = \frac{\sigma^2_{общ}}{\sigma^2 ост}$$

$$\sigma^2_{факт} > \sigma^2_{ост}$$

Любая сумма квадратов отклонений связана с числом степеней свободы df, то есть с числом свободы независимого варьирования признака, который определяется размером выборки и числом определяемых по ней констант.

$$\sum (y - \bar{y})^2 = \sum (\hat{y}_x - \bar{y})^2 + \sum (y - \hat{y}_x)^2$$

- Степени свободы:

$$(n-1) = 1 + (n-2)$$

- Дисперсия на 1 степень свободы:

$$D_{общ} = \frac{\sum (y - \bar{y})^2}{n-1}$$

$$D_{факт} = \frac{\sum (\hat{y}_x - \bar{y})^2}{1}$$

$$D_{ост} = \frac{\sum (y - \hat{y}_x)^2}{n-2}$$

F-статистика:

$$F = \frac{D_{факт}}{D_{ост}}$$

- Проверка:
- 1. $H_0 : r^2 = 0; H_1 : r^2 \neq 0$
- 2. $\alpha = 0{,}05$
- 3. F-статистика
- 4. $F_{кр(n-2,\alpha)}$
- 5. $F = \frac{r^2}{1-r^2} * (n-2)$

1. $H_0 : D_{факт} = D_{ост}; H_1 : D_{факт} > D_{ост}$

2. $$m_r = \sqrt{\frac{1-r^2}{n-2}}$$ – стандартная ошибка.

3. $$t_r = \frac{r}{\sqrt{1-r^2}} * \sqrt{n-2}$$ –(*)

4. $$t^2{}_r = F$$

Формула (*) рекомендуется при большом *n* и если *r* не стремится к 1 или к -1. Если же , то распределение его оценок отличается от нормального распределения или распределения Стьюдента.
Чтобы избежать этого затруднения Фишером было предложено ввести вспомогательную величину Z:

$$Z = \frac{1}{2} * \ln \frac{1+r}{1-r}$$

Если $-1 < r < +1$ то $(-\infty) < Z < (+\infty)$

Тогда стандартная ошибка для Z: $m_z = \frac{1}{\sqrt{n-3}}$

Тогда: $H_0 : r = 0; H_1 : R \neq 0$

$$\frac{Z}{m_z} = t_z$$

Существуют таблицы для оценки значимости по этим формулам.