

1. Нужно построить такой набор точек  $(X_i, Y_i)$ , чтобы:

в первой подвыборке  $i=1..50$  регрессия  $Y_i$  на  $X_i$  имела положительный наклон  $\beta_1 > 0$ ;

во второй подвыборке  $i=51..100$  тоже положительный наклон  $\alpha_1 > 0$ ;

но на всей выборке  $i=1..100$  наклон оказался отрицательным  $\gamma_1 < 0$ .

Это возможно, если данные состоят из двух разных групп. Внутри каждой группы  $Y$  растёт с  $X$ , но сами группы расположены так, что у второй группы  $X$  в среднем больше, а  $Y$  в среднем меньше

То есть на диаграмме рассеяния должны быть:

левое облако точек (первая группа): малые  $X$ , большие  $Y$ , слабый положительный наклон;

правое облако точек (вторая группа): большие  $X$ , меньшие  $Y$ , тоже положительный наклон;

если провести одну прямую по всем точкам, она будет убывающей из-за различия средних уровней групп

```
In [2]: import numpy as np
import matplotlib.pyplot as plt

np.random.seed(7)

n1 = n2 = 50

# группа 1
x1 = np.random.normal(loc=2.0, scale=0.6, size=n1)
y1 = 70 + 3.0*x1 + np.random.normal(scale=2.0, size=n1)

# группа 2
x2 = np.random.normal(loc=7.0, scale=0.7, size=n2)
y2 = 35 + 4.5*x2 + np.random.normal(scale=2.5, size=n2)

# выборка
x = np.concatenate([x1, x2])
y = np.concatenate([y1, y2])

# линейные регрессии
b1, b0 = np.polyfit(x1, y1, 1)
a1, a0 = np.polyfit(x2, y2, 1)
g1, g0 = np.polyfit(x, y, 1)

print("Оцененные наклоны:")
print("β1 =", b1)
```

```
print("α1 =", a1)
print("γ1 =", g1)
```

Оцененные наклоны:

$\beta_1 = 4.114890988943115$

$\alpha_1 = 4.667067210001288$

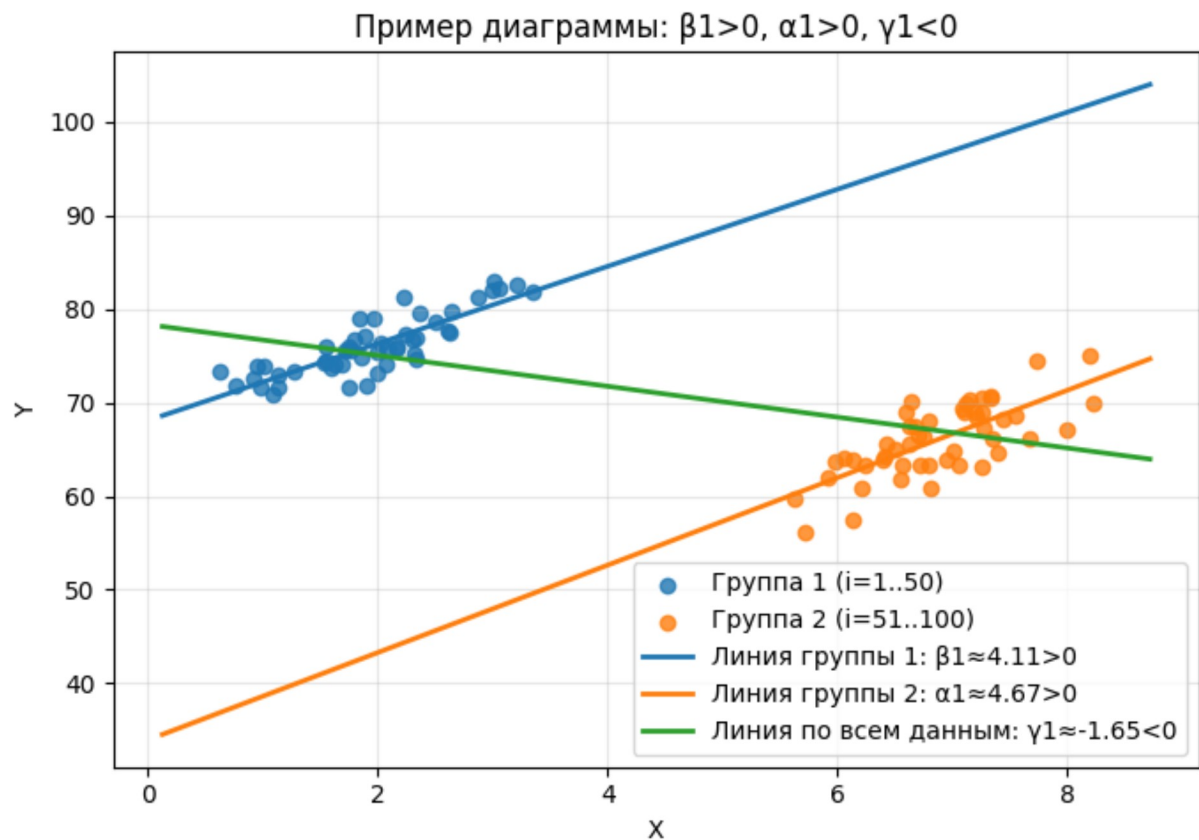
$\gamma_1 = -1.6477698033023258$

Отлично, теперь построю сам график

```
In [3]: plt.figure(figsize=(7,5))
plt.scatter(x1, y1, alpha=0.8, label="Группа 1 (i=1..50)")
plt.scatter(x2, y2, alpha=0.8, label="Группа 2 (i=51..100)")

xx = np.linspace(min(x)-0.5, max(x)+0.5, 200)
plt.plot(xx, b0 + b1*xx, linewidth=2, label=f"Линия группы 1:  $\beta_1 \approx \{b1:.2f\} > 0$ ")
plt.plot(xx, a0 + a1*xx, linewidth=2, label=f"Линия группы 2:  $\alpha_1 \approx \{a1:.2f\} > 0$ ")
plt.plot(xx, g0 + g1*xx, linewidth=2, label=f"Линия по всем данным:  $\gamma_1 \approx \{g1:.2f\} < 0$ ")

plt.xlabel("X")
plt.ylabel("Y")
plt.title("Пример диаграммы:  $\beta_1 > 0$ ,  $\alpha_1 > 0$ ,  $\gamma_1 < 0$ ")
plt.legend()
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()
```



2. Реальный пример: X — количество часов подготовки к экзамену.

$Y$  — итоговый балл экзамена.

Полная выборка из 100 наблюдений - это все студенты, но они делятся на две группы: подвыборка 1 ( $i=1..50$ ): сильные студенты.

Обычно они учатся меньше ( $X$  меньше), но из-за хорошей базы получают высокий балл ( $Y$  больше).

Среди сильных: кто готовился больше - получает чуть лучше, поэтому  $\beta_1 > 0$ .  
подвыборка 2 ( $i=51..100$ ): слабые студенты.

Им приходится учиться больше ( $X$  больше), но результат всё равно в среднем ниже ( $Y$  меньше)

Среди слабых: дополнительные часы реально улучшают результат, поэтому  $\alpha_1 > 0$ .

На всей выборке  $\gamma_1 < 0$ , потому что при переходе от сильных к слабым  $X$  в среднем растёт, а  $Y$  в среднем падает.

То есть смешение двух групп с разными средними уровнями переворачивает общий знак зависимости

3. Теперь дополнительно дано:  $\alpha_1 > \beta_1$ .

В терминах примера это означает, что отдача от подготовки у слабых студентов выше, чем у сильных.

Почему так бывает:

сильные и так близки к максимуму, у них эффект потолка: дополнительные часы дают небольшой прирост;

слабые закрывают пробелы, поэтому каждый новый час добавляет заметнее — наклон у них выше.

Запишу уравнение регрессии, которое описывает обе группы одновременно

Введу индикатор группы:

$D_i = 0$  для сильных ( $i=1..50$ ),  $D_i = 1$  для слабых ( $i=51..100$ ).

Тогда общая модель:

$$Y_i = b_0 + b_1 X_i + b_2 D_i + b_3 (D_i X_i) + \text{ошибка}_i$$

Объясню:

если  $D_i=0$  (сильные):  $Y_i = b_0 + b_1 X_i$ , значит  $b_1 = \beta_1 > 0$ .

если  $D_i=1$  (слабые):  $Y_i = (b_0 + b_2) + (b_1 + b_3) \cdot X_i$ , значит наклон  $b_1 + b_3 = \alpha_1$ . Условие  $\alpha_1 > \beta_1$  означает  $b_3 > 0$ .

А общий наклон по всей выборке может стать отрицательным из-за того, что  $b_2$  отрицателен (слабые в среднем ниже по  $Y$ ) и группы сильно разнесены по  $X$