

```
In [3]: from statsmodels.api import OLS, add_constant
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

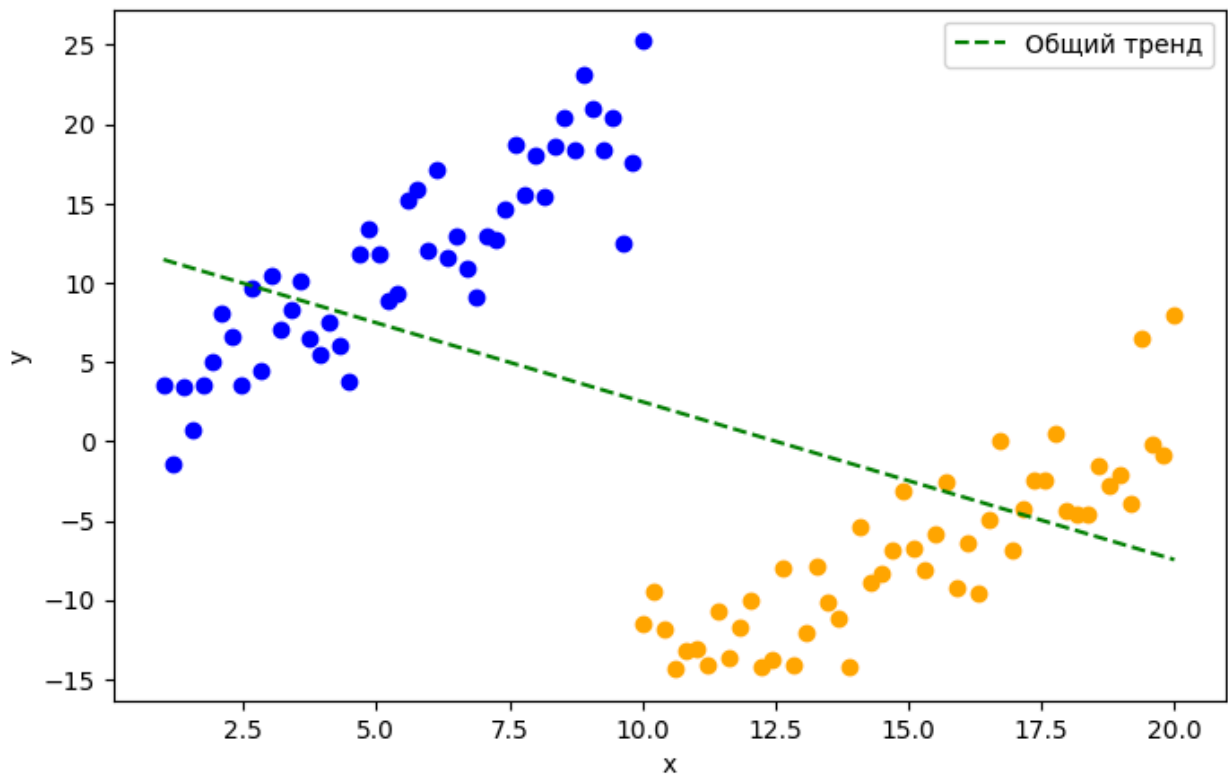
```
In [4]: np.random.seed(52)

X1 = np.linspace(1, 10, 50)
Y1 = 2 * X1 + np.random.normal(0, 3, 50)

X2 = np.linspace(10, 20, 50)
Y2 = 1.5 * X2 - 30 + np.random.normal(0, 3, 50)
X = np.concatenate([X1, X2])
Y = np.concatenate([Y1, Y2])

plt.figure(figsize=(8, 5))
plt.scatter(X1, Y1, c="blue")

plt.scatter(X2, Y2, c="orange")
plt.plot(X, np.polyval(np.polyfit(X, Y, deg=1), X), "g--", label = "Общий тренд")
plt.xlabel("x")
plt.ylabel("y")
plt.legend()
plt.show()
```



Пусть  $x$  - возраст, а  $y$  - средний доход. В первой подвыборке люди от 20 до 35 лет, а во второй от 35 до 45. В обеих группах доход растёт не пропорционально возрасту, но у старых людей его меньше, так как к

перспективе молодых сотрудников приглядываются работодатели больше, нежели чем к взрослым людям

Это возможно, если во второй группе связь дохода и возраста выражена сильнее, например быстрее растет доход, несмотря на то что в целом он стал меньше. Пусть в старшей группе причиной роста дохода становится накопленный опыт.  $Y_i = \text{sig\_0} + \text{sig\_1} * X_i + \text{sig\_2} * d_i + \text{sig\_3} * d_i * d_x + \text{epsilon}$   $d_i$  - идентификатор принадлежности к какой-то группе, он равен либо 0 либо 1  $\text{sig\_0}$  - свободный член  $\text{sig\_1}$  - коэффициент оценивающий вклад  $X$   $\text{sig\_2}$  - коэффициент при  $d_i$   $\text{sig\_3}$  - коэффициент отражающий различие в наклоне между двумя группами  $\text{epsilon}$  - погрешность