

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

import statsmodels.api as sm
import scipy.stats as stats

from sklearn.metrics import r2_score, mean_absolute_percentage_error
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

sns.set()
pd.options.display.max_columns = None

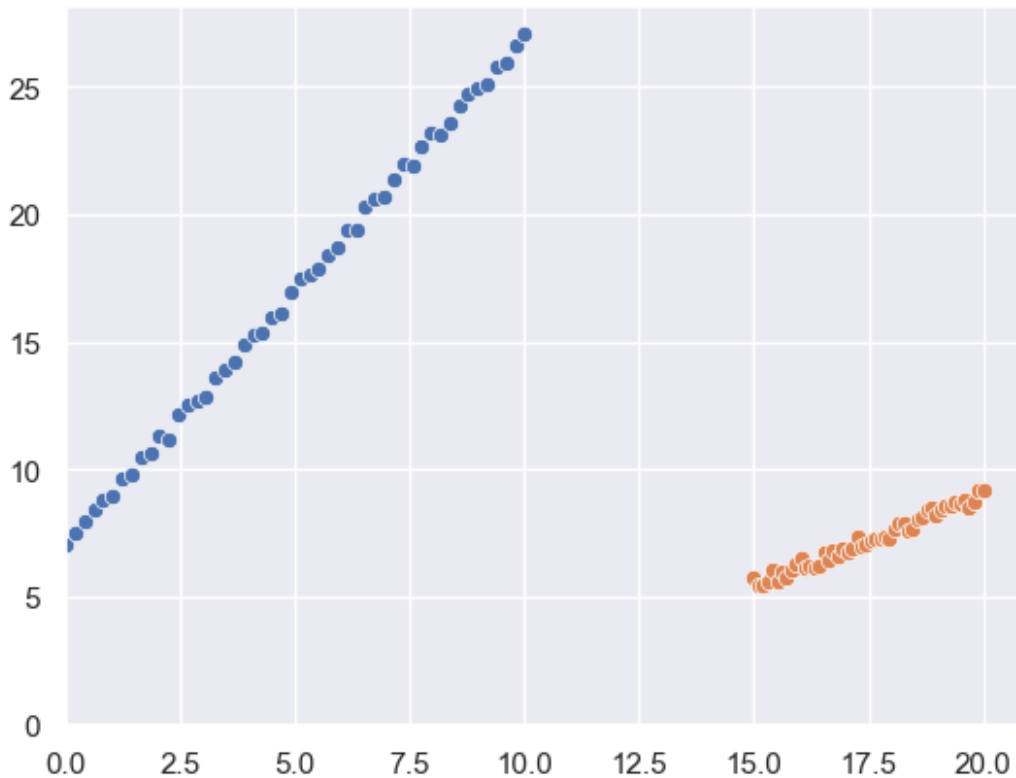
def f1(x):
    return 2 * x + 7 + np.random.uniform(-0.3, 0.3, (1, 1))[0][0]

def f2(x):
    return 0.7 * x + -5 + np.random.uniform(-0.3, 0.3, (1, 1))[0][0]

x1 = np.linspace(0, 10, 50)
x2 = np.linspace(15, 20, 50)
y1 = list(map(f1, x1))
y2 = list(map(f2, x2))

sns.scatterplot(x=x1, y=y1)
sns.scatterplot(x=x2, y=y2)
plt.xlim(0)
plt.ylim(0)

(0.0, 28.135057134410385)
```



Выше пример такого графика

2

Даша могла получить такую зависимость в случае, если например первые 50 и вторые 50 наблюдений различаются по какой-то третьей переменной. Пример: зависимость уровня стресса от нагрузки у мужчин и женщин для выборок из мужчин и женщин отдельно возможно возникнет положительная линейная взаимосвязь, для объединения из-за смещения может возникнуть отрицательная взаимосвязь

3

Такая зависимость может быть обусловлена опять же третьим фактором, различия a_1 и b_1 могут свидетельствовать о том, что под влиянием третьего фактора не просто выше уровень, а более сильное влияние. Чтобы учесть влияние третьего фактора можно ввести этот случайные фактор в регрессию

пусть $p = 1$, если $i > 50$, 0 иначе
Тогда регрессия может иметь вид:

$$y = b_1 * x + b_0 + b_2 * p + b_3 * p *$$