Лабораторная работа 1

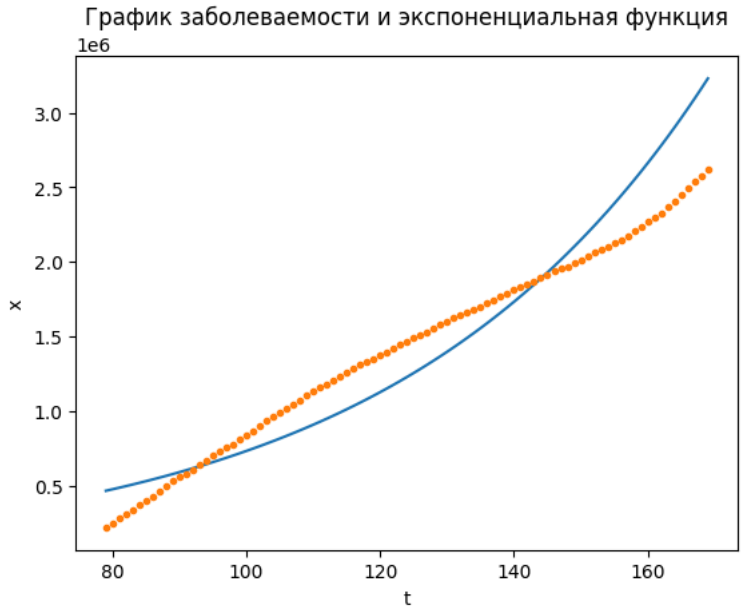
Тема: О росте больных ковидом

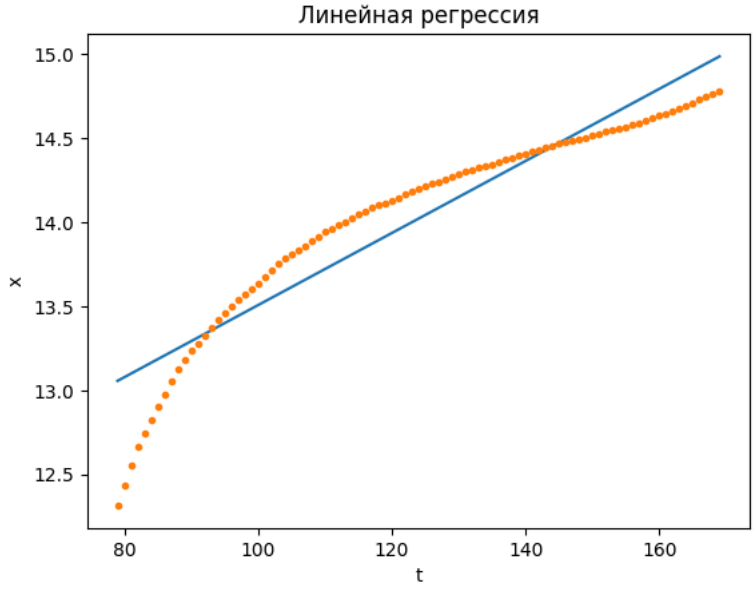
Вариант 6 (сдвиг на 2 месяца)

Возьмите открытые данные о росте числа зараженных COVID-19 в мире за три месяца. Первый месяц из трех определяется сдвигом на номер по списку с февраля 2020 года. Проверьте гипотезу о том, что этот рост описывается экспоненциальной функцией. Полученное p-значение сравните с уровнем значимости 0,05.

```
import pandas as pd
import numpy as np
import datetime
import scipy
import matplotlib.pyplot as plt

bins = 8

df = pd.read_csv('national-history.csv', sep=',')
df=df.iloc[::-1].reset_index(drop=True)
df['date'] = pd.to_datetime(df['date'], format='%Y-%m-%d')
df = df.loc[(df['date'] >= '2020-04-01')
                    & (df['date'] < '2020-07-01')]
df = df[["date", "positive"]]
df["y"] = np.log(df["positive"])

b, a = np.polyfit(df.index.values, df["y"], 1)

K, p = scipy.stats.pearsonr(df["y"],
  [b * index + a for index in df.index.values])

plt.title("График заболеваемости и экспоненциальная функция")
plt.xlabel("t")
plt.ylabel("x")

x = df.index.values

plt.plot(x, np.exp(b * x + a), x, df["positive"], ".")

plt.show()

plt.title("Линейная регрессия")
plt.xlabel("t")
plt.ylabel("x")

x = df.index.values

plt.plot(x, b * x + a, x, df["y"], ".")

plt.show()


print("K =", K, "p =", p)

print("Оценка адекватности:")

df["e"] = df["y"] - (b * df.index.values + a)
df["e_disp_i"] = (df["e"] - df["e"].mean()) ** 2

dispersion = df["e_disp_i"].mean() ** 0.5

min_e = min(df["e"]) - 0.001
max_e = max(df["e"])

df['possible'] = pd.cut(df["e"], [min_e + i * (max_e - min_e) / bins for i in range(0, bins + 1)],
                        labels=[i for i in range(0, bins)])

emps = [df[df["possible"] == i].count()["possible"] for i in range(0, bins)]

exps = [(((scipy.stats.norm.cdf((min_e + (i + 1) * (max_e - min_e) / bins) / dispersion) -
        scipy.stats.norm.cdf((min_e + i * (max_e - min_e) / bins) / dispersion)) * df.count())["possible"])
        for i in range(0, bins)]

emps[-1] += sum(exps) - sum(emps)

G = 1.98

K, p = scipy.stats.chisquare(f_obs=emps, f_exp=exps, ddof=1)
```

```
print("K =", K, "p =", p)

if K > G:
  print(f"Гипотеза отвергается, K > G ({K} > {G})")
else:
  print(f"Гипотеза подтвердилась, K < G ({K} < {G})")
```






```
K = 0.9419274856635396 p = 5.949288630770426e-44
Оценка адекватности:
K = 55.093831013694675 p = 4.4375941829891163e-10
Гипотеза отвергается, K > G (55.093831013694675 > 1.98)
```

K близок к 1, что говорит о сильной линейной зависимости. Кроме того значение p крайне мало. Очень маленькое p-значение чуть меньше $6 \cdot 10^{-44}$ – это вероятность случайной величины, распределенной по Стьюденту с 91 - 2 = 89 степенями свободы принять значение по модулю больше $T \approx 1{,}98$.

Однако оценка адекватности провалилась. K > 1.98, следовательно гипотеза отвергается и можно сделать вывод, что если рост заболеваемых сначала и был экспоненциальным, на выбранный период он таковым не является.