Домашнє завдання 3. Багатофакторні регресивні моделі та моделювання нецінових факторів впливу на криву попиту.

Виконав: Микола Трохимович

Завдання 1.

Побудова лінії лінійного тренду. У вас є історичні дані про найкращі результати з бігу на 100 м на Олімпійських іграх для чоловіків і жінок. Побудуйте лінію лінійного тренду впливу року проведення Олімпійських ігор на результат, окремо для чоловіків і окремо для жінок, використовуючи в Excel модель:

$$Time = \alpha_0 + \alpha_1 \times YEAR + \epsilon$$

1.1 Знайдіть значення α_0, α_1 та значення R^2.

1.2 Наскільки доброю є модель в цілому на основі показника R^2?

1.3 Чи порадили б ви використовувати лінійну регресивну модель для передбачення найкращого часу у бігу на 100 м для чоловіків чи жінок? Відповідь обґрунтуйте.

1.4 У чому різниця між пояснюючою та прогностичною силою моделі?

Для даного завдання використаю Python. Вхідні дані згідно умов задачі.

Розв'язок:

In [152]:

```
import pandas as pd
import numpy as np
from IPython.core.display import HTML

def excel2python(x):
    return float(x.replace(',','.'))

display(HTML(f'<h3>Вхідні дані</h3>'))
data = pd.read_csv('data/task1.csv', skiprows = 2, names = ['year', 'man', 'woman'])
display(data)
```

Вхідні дані

	year	man	woman
0	1896	11,80	NaN
1	1900	10,8	NaN
2	1904	10,8	NaN
3	1908	10,8	NaN
4	1912	10,8	NaN
5	1920	10,8	NaN
6	1924	10,6	NaN
7	1928	10,8	12,2
8	1932	10,38	11,9
9	1936	10,3	11,5
10	1948	10,3	11,9
11	1952	10,4	11,5
12	1956	10,5	11,5
13	1960	10,2	11
14	1964	10	11,4
15	1968	9,95	11
16	1972	10,14	11,07
17	1976	10,06	11,08
18	1980	10,25	11,06
19	1984	9,99	10,97
20	1988	9,92	10,54
21	1992	9,96	10,82
22	1996	9,84	10,94
23	2000	9,87	11,12
24	2004	9,85	10,93
25	2008	9,69	10,78
26	2012	9,63	10,75
27	2016	9,81	10,71

1.1 Знайдіть значення α_0, α_1 та значення R.

$$Time = \alpha_0 + \alpha_1 \times YEAR + \epsilon$$

Для чоловіків:

In [160]:

```
import statsmodels.api as sm
from statsmodels.sandbox.regression.predstd import wls_prediction_std

X = sm.add_constant(data.year.values)
y = data.man.apply(excel2python).values.astype(float)

model = sm.OLS(y, X)
results1 = model.fit()
print(results1.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Parameters: alpha_0 = {results1.params.round(4)[0]}, alpha_1 = {results.params.round(4)[1]}')
print('R2: ', results1.rsquared.round(4))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.829
Model:                            OLS   Adj. R-squared:                  0.822
Method:                 Least Squares   F-statistic:                     125.8
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           1.85e-11
Time:                        18:20:02   Log-Likelihood:                 6.0316
No. Observations:                  28   AIC:                            -8.063
Df Residuals:                      26   BIC:                            -5.399
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         33.2501      2.047     16.241      0.000      29.042      37.458
x1            -0.0117      0.001    -11.215      0.000      -0.014      -0.010
==============================================================================
Omnibus:                       28.659   Durbin-Watson:                   1.447
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               66.105
Skew:                           2.046   Prob(JB):                     4.42e-15
Kurtosis:                       9.318   Cond. No.                     1.05e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.05e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Відповідь:

```
Parameters: alpha_0 = 33.2501, alpha_1 = -0.0142
R2:  0.8287
```

Для жінок:

In [40]:

```
import statsmodels.api as sm
from statsmodels.sandbox.regression.predstd import wls_prediction_std

X2 = sm.add_constant(data.dropna().year.values)
y2 = data.dropna().woman.apply(excel2python).values.astype(float)

model = sm.OLS(y2, X2)
results2 = model.fit()
print(results2.summary())

print()
print()
display(HTML(f'<h2>Відповідь:</h2>'))
print(f'Parameters: alpha_0 = {results2.params.round(4)[0]}, alpha_1 = {results.params.round(4)[1]}')
print('R2: ', results2.rsquared.round(4))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.738
Method:                 Least Squares   F-statistic:                     57.40
Date:                Tue, 06 Oct 2020   Prob (F-statistic):           3.72e-07
Time:                        21:25:13   Log-Likelihood:                2.7467
No. Observations:                  21   AIC:                            -1.493
Df Residuals:                      19   BIC:                            0.5957
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         39.1882      3.698     10.598      0.000      31.449      46.928
x1            -0.0142      0.002     -7.577      0.000      -0.018      -0.010
==============================================================================
Omnibus:                        0.448   Durbin-Watson:                   1.623
Prob(Omnibus):                  0.799   Jarque-Bera (JB):                0.425
Skew:                          -0.293   Prob(JB):                        0.808
Kurtosis:                       2.623   Cond. No.                     1.50e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.5e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Відповідь:

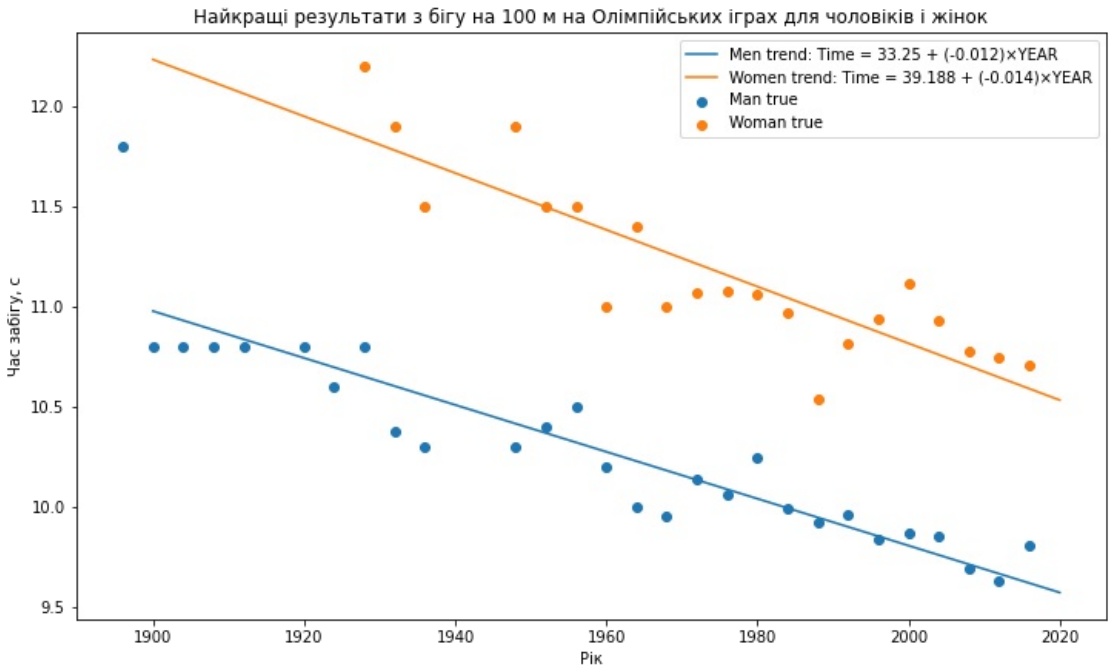
```
Parameters: alpha_0 = 39.1882, alpha_1 = -0.0142
R2:  0.7513
```

In [119]:

```
import matplotlib.pyplot as plt
from pylab import rcParams
import copy
rcParams['figure.figsize'] = 12, 7
x = np.linspace(1900, 2020, 1000)

def line(x,params):
    return params[0] + x*params[1]

def print_line(params):
    p = copy.deepcopy(params)
    p = p.round(3)
    return f'Time = {p[0]} + ({p[1]})×YEAR'


plt.scatter(X[:,1],y, label = 'Man true')
plt.plot(x, line(x,results1.params), label='Men trend: ' + print_line(results1.params))

plt.scatter(X2[:,1],y2, label = 'Woman true')
plt.plot(x, line(x,results2.params), label='Women trend: '+  print_line(results2.params))

plt.xlabel('Рік')
plt.ylabel('Час забігу, с')
plt.title('Найкращі результати з бігу на 100 м на Олімпійських іграх для чоловіків і жінок')
plt.legend()
plt.plot()
```

Out[119]:

```
[]
```



1.2 Наскільки доброю є модель в цілому на основі показника R_2 ?

Відповідь: R_2 для моделі для *чоловіків* становить ~83%, R_2 для моделі для *жінок* становить ~75%. Якщо оцінювати вцілому то обидві моделі можна назвати допустимими, а модель для чоловіків навіть достатньо хорошою. Даний показник коливається від 0 до 1, де 1 означає повну функціональну залежність (ідеальний варіант). Обидві моделі достаньо не погано описують функціональну залежність, що ми досліджуємо.

1.3 Чи порадили б ви використовувати лінійну регресивну модель для передбачення найкращого часу у бігу на 100 м для чоловіків чи жінок? Відповідь обґрунтуйте.

Якщо мова йде про передбачення найкращого часу у бігу на 100 м у досліджуваний період (1886-2016рр.), так би мовити інтерполяцію, тоді так, дана модель є достатньо непоганою. Її можна використовувати наприклад для оцінки швидкості на чемпіонаті в періоді між олімпійськими іграми цього періоду. Також, скоріше за все модель буде непогана для невеликого горизонту подій (на декілька ігор вперед).

Проте якщо ми будемо передбачати результати Олімпійських ігор що стануться через 300 років, тоді дана модель буде не зовсім коректна. Ми розуміємо, що здатності людини є обмежені, і спадання функції буде сповільнюватися з часом. Тому я б краще використав якусь нелінійну залежність, що мала б спадний характер і з часом це спадання сповільнювалося б.

1.4 У чому різниця між пояснюючою та прогностичною силою моделі?

Прогностична сила визначає частку інформації в цільової змінної, яку можуть описати інші змінні. Фактично це міра того наскільки модель є точною.

Пояснююча сила показує наскільки добре модель описує реальні процеси, що відбуваються в системі, наскільки якісно властивості моделі відповідають властивостям модельованого об'єкта.

У нашому контретному випадку можемо бачити, що модель є достатньо точною, тобто має непогану прогностичну силу. Проте пояснююча сила є дещо гіршою так як не всі властивості реального світу знаходять відповідники в моделі (обмеженість людського ресурсу).

Завдання 2. Побудова лінії тренду. Розгляньте дані зараження людей коронавірусом covid 19.

2.1 Побудуйте графічне зображення кількості заражених людей по днях, а також лінійного лінію тренду, використавши додатково функціонального виразу для цього тренду і коефіцієнт детермінації R^2.

2.2 Побудуйте графічне зображення кількості заражених людей по днях, а також експоненціальну лінію тренду, використавши додатково функціонального виразу для цього тренду і коефіцієнт детермінації R^2 .

2.3 Чи рекомендували б ви використовувати експоненціальний або лінійний тренд в якості моделі для прогнозування захворюваності на covid-19?

2.4 Який клас гнучких функцій ви б запропонували для моделювання поширення захворюваності на covid-19?

2.5 Зробіть економетричну оцінку динаміки захворюваності на covid-19 для запропонованого вами класу функцій з п. 2.4.

Для початку спарсимо дані про хворих на ковід. Для цього використаємо сайт РНБО України https://covid19.rnbo.gov.ua/ (https://covid19.rnbo.gov.ua/). Тут доступні всі актуальні дані зі всього світу. Але для нашого дослідження використаємо тільки дані для України. Використавши маленьку хитрість напишемо скрипт для отримання даних в сирому вигляді, придатному до аналізу і моделювання.

Парсинг даних з сайту РНБО

In [120]:

```
from datetime import date, timedelta
from tqdm.auto import tqdm
import pandas as pd
import requests

request_template = "https://api-covid19.rnbo.gov.ua/data?to="
data = []

for i in tqdm(range(220)):
    today = date.today() - timedelta(days=i)

    # dd-mm-YY
    d1 = today.strftime("%Y-%m-%d")
    response = requests.get(request_template+d1)
    tmp = response.json()['world']
    for c in tmp:
        c['date'] = d1
        c['day of observation'] = 220-i
    data += tmp
```

Залишаємо тільки дані для України

In [135]:

```
df = pd.DataFrame(data)
df_ukr = df[df.country=='Ukraine'].sort_values('date').reset_index(drop = True)
df_ukr['day of observation'] = df_ukr.index.values
df_ukr.date = pd.to_datetime(df_ukr.date)
df_ukr.head()
```

Out[135]:

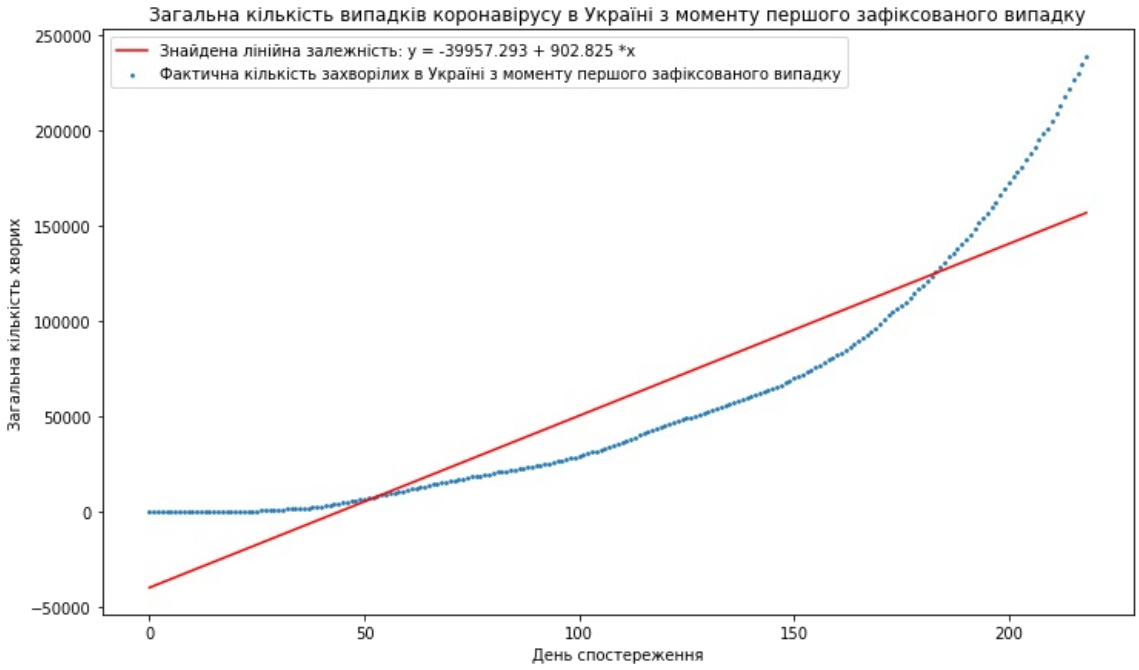
	id	label	country	confirmed	deaths	recovered	existing	suspicion	lat	lng	delta_confirmed	delta_deaths	delta
0	4905	{'en': 'Ukraine', 'uk': 'Україна'}	Ukraine	1	0	0	1	0	49.288113	30.336351	0	0	
1	4905	{'en': 'Ukraine', 'uk': 'Україна'}	Ukraine	1	0	0	1	0	49.288113	30.336351	0	0	
2	4905	{'en': 'Ukraine', 'uk': 'Україна'}	Ukraine	1	0	0	1	0	49.288113	30.336351	0	0	
3	4905	{'en': 'Ukraine', 'uk': 'Україна'}	Ukraine	1	0	0	1	0	49.288113	30.336351	0	0	
4	4905	{'en': 'Ukraine', 'uk': 'Україна'}	Ukraine	1	0	0	1	0	49.288113	30.336351	0	0	

2.1 Побудуйте графічне зображення кількості заражених людей по днях, а також лінійну лінію тренду, використавши додатково функціонального виразу для цього тренду і коефіцієнт детермінації R^2

Варто зазначити, що використовуючи середовище Python ми стаємо набагато гнучнішими до досліджень, адже не обмежуємося моделями які наявні в пакетних рішеннях та можемо аналізувати значно більші об'єми інформації.

In [171]:

```
import matplotlib.pyplot as plt
from scipy.optimize import curve_fit
from sklearn.metrics import r2_score

# function for trend fitting
def func(x, a, b):
    return a + b*x

# data to observe
x = df_ukr['day of observation'].values
y = df_ukr.confirmed

# plot the actual data
plt.scatter(x, y, label='Фактична кількість захворілих в Україні з моменту першого зафіксованого випадку', s=3)

# finding the trend
popt, pcov = curve_fit(func, x, y)

#plotting the trend
plt.plot(x, func(x, *popt), 'r-',
         label='Знайдена лінійна залежність: y = %5.3f + %5.3f *x' % tuple(popt))

plt.xlabel('День спостереження')
plt.ylabel('Загальна кількість хворих')
plt.title('Загальна кількість випадків коронавірусу в Україні з моменту першого зафіксованого випадку')
plt.legend()
plt.show()


def line(x,params):
    return params[0] + x*params[1]

y_true = y
y_pred = func(x, *popt)

r2 = r2_score(y_true, y_pred)

display(HTML(f'<h4>Коефіцієнт детермінації R2 = {r2}</h4>'))
```

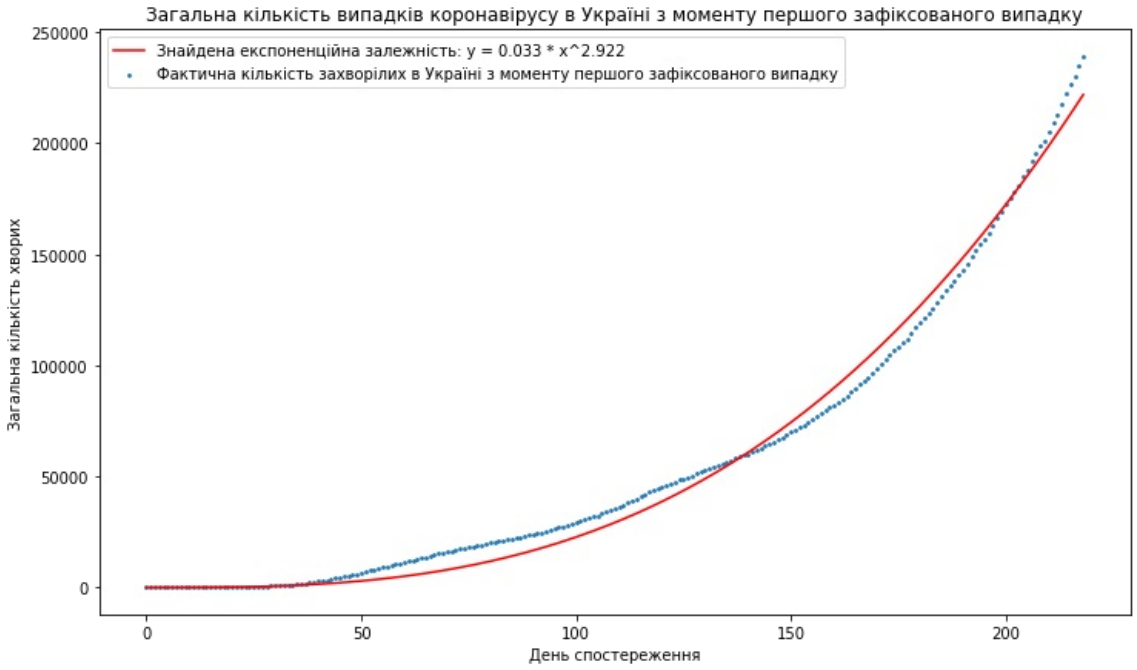


Коефіцієнт детермінації R2 = 0.8378727296346732

2.2 Побудуйте графічне зображення кількості заражених людей по днях, а також експоненціальну лінію тренду, використавши додатково функціонального виразу для цього тренду і коефіцієнт детермінації R^2

In [211]:

```
import matplotlib.pyplot as plt
from scipy.optimize import curve_fit
from sklearn.metrics import r2_score

# function for trend fitting
def func(x, a, b):
    return a*x**b

# data to observe
x = df_ukr['day of observation'].values
y = df_ukr.confirmed

# plot the actual data
plt.scatter(x, y, label='Фактична кількість захворілих в Україні з моменту першого зафіксованого випадку', s=3)

# finding the trend
popt, pcov = curve_fit(func, x, y)

#plotting the trend
plt.plot(x, func(x, *popt), 'r-',
         label='Знайдена експоненційна залежність: y = %5.3f * x^%5.3f' % tuple(popt))

plt.xlabel('День спостереження')
plt.ylabel('Загальна кількість хворих')
plt.title('Загальна кількість випадків коронавірусу в Україні з моменту першого зафіксованого випадку')
plt.legend()
plt.show()


def line(x,params):
    return params[0] + x*params[1]

y_true = y
y_pred = func(x, *popt)

r2 = r2_score(y_true, y_pred)

display(HTML(f'<h4>Коефіцієнт детермінації R2 = {r2}</h4>'))
```



Коефіцієнт детермінації R2 = 0.9909529658555227

2.3 Чи рекомендували б ви використовувати експоненціальний або лінійний тренд в якості моделі для прогнозування захворюваності на covid-19?

Очевидно, що для прогнозування захворюваності я б рекомендував краще використовувати експоненційну залежність. Як ми можемо бачити з графіку вона значно краще апроксимує кількість хворих по днях. Також коефіцієнт детермінації R^2 становить 0.99, що означає майже повну віповідність. Тоді якби це була 1 мали б функціональну залежність. Цей показник значно кращий ніж відповідник за лінійної залежності.

2.4 Який клас гнучких функцій ви б запропонували для моделювання поширення захворюваності на covid-19?

Для моделювання поширення захворюваності я б пропонував використовувати клас логістичних функцій вигляду $y = \frac{a}{(1+b*e^{c*x})}$. Даний клас функцій найбільш вдало описує природу епідемії, а саме спочатку має плавне зростання, що з часом наростає, і перетворюється в різке зростання. Це моделює поведінку епідемії, коли мало хворих - вони заражають мало людей, коли ж кількість хворих зростає, то і росте кількість нових випадків. Потім спостерігається точка перегину, коли з кожним днем захворюваність починає бути меншою, і потім функція виходить на плато. Це відбувається через те, що починають вживатися заходи або ж вже банально немає кому хворіти, так як людей є обмежена кількість. Відповідно так як людей є обмежена кількість то і наша модель має бути обмеженою зверху. Таку поведінку має логістичний тренд. Він обмежений зверху. Через логіку описану вище, я б рекомендував використовувати саме такий тренд.

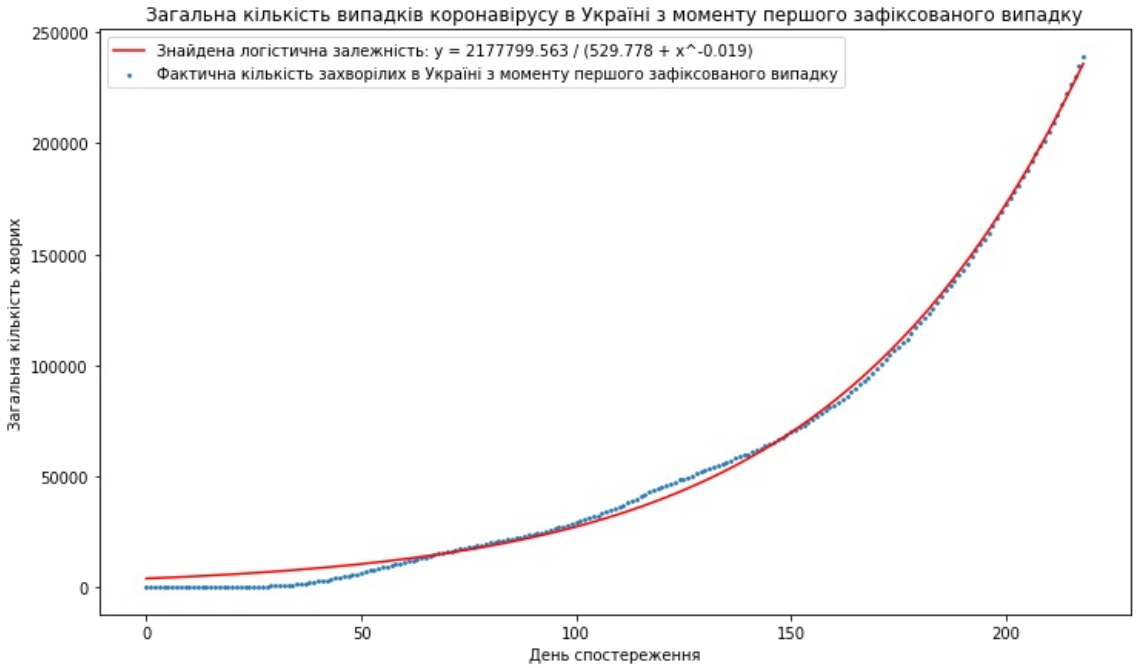
2.5 Зробіть економетричну оцінку динаміки захворюваності на covid-19 для запропонованого вами класу функцій з п. 2.4.

Як можемо бачити з подальшого графіку і коефіцієнту детермінації, даний клас функцій дуже добре описує поведінку даних. Маємо коефіцієнт детермінації 0.997, що вище ніж у експоненційного класу функцій. Також даний клас є ближчим до природи даних, та відповідає властивостям притаманним епідемілогічним даним.

In [219]:

```
import matplotlib.pyplot as plt
from scipy.optimize import curve_fit
from sklearn.metrics import r2_score

# function for trend fitting
def func(x, a, b,c):
    return a/(1+b*np.exp(x*c))


# data to observe
x = df_ukr['day of observation'].values
y = df_ukr.confirmed

# plot the actual data
plt.scatter(x, y, label='Фактична кількість захворілих в Україні з моменту першого зафіксованого випадку', s=3)

# finding the trend
popt, pcov = curve_fit(func, x, y)

#plotting the trend
plt.plot(x, func(x, *popt), 'r-',
         label='Знайдена логістична залежність: y = %5.3f / (%5.3f + x^%5.3f)' % tuple(popt))

plt.xlabel('День спостереження')
plt.ylabel('Загальна кількість хворих')
plt.title('Загальна кількість випадків коронавірусу в Україні з моменту першого зафіксованого випадку')
plt.legend()
plt.show()


def line(x,params):
    return params[0] + x*params[1]

y_true = y
y_pred = func(x, *popt)

r2 = r2_score(y_true, y_pred)

display(HTML(f'<h4>Коефіцієнт детермінації R2 = {r2}</h4>'))
```

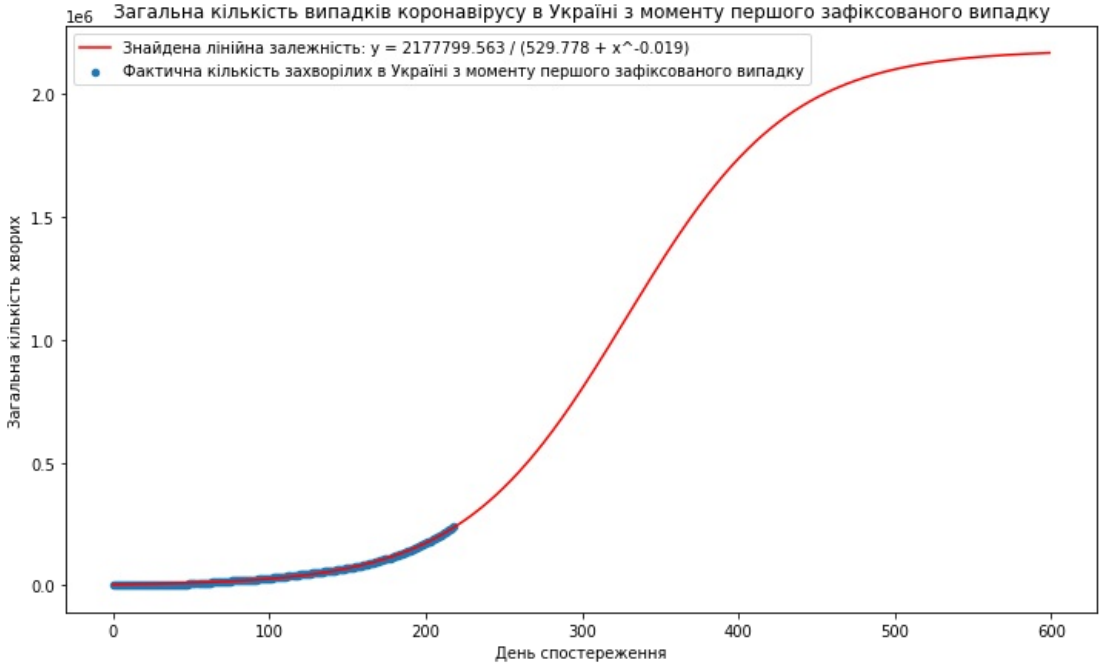
```
/Users/trokhymovych/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:7: RuntimeWarning: o
verflow encountered in exp
  import sys
/Users/trokhymovych/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:7: RuntimeWarning: o
verflow encountered in multiply
  import sys
```



Коефіцієнт детермінації R2 = 0.9969910356068628

In [220]:

```
x2 = np.arange(600)
# plot the actual data
plt.scatter(x, y, label='Фактична кількість захворілих в Україні з моменту першого зафіксованого випадку', s=20)
plt.plot(x2, func(x2, *popt), 'r-',
         label='Знайдена лінійна залежність: y = %5.3f / (%5.3f + x^%5.3f)' % tuple(popt))
plt.xlabel('День спостереження')
plt.ylabel('Загальна кількість хворих')
plt.title('Загальна кількість випадків коронавірусу в Україні з моменту першого зафіксованого випадку')
plt.legend()
plt.show()
```



Також для себе зробив невеликий додатковий аналіз і прогнозування (дуже грубе і можливо не точне). Колистуючись моделлю логістичного тренду можемо отримати такі результати з розвитку епідемії в Україні

Оцінка захворюваності на кінець епідемії ~2.1 млн. людей.

На кінець року цифра досягне 830000 людей (+594000 від сьогодні)

Хоча для прогнозування я б тут іншу модель використовував, з вікном, що приблизно відповідає інкубаційному періоду (7-14 днів)

Завдання 3

Розглянемо базу даних з продажами протягом місяця деякої мережі, а також фактори, що на нашу думку впливають на продажі:

1) площа магазину –area;

2) кластер магазину, визначений історично -cluster; (0,...,5)

3) наявність міні пекарні у магазині –bread; (0,1)

4) наявність паркінгу біля магазину –parking; (0,1)

Нехай sales є залежною змінною у цій лінійній регресивній моделі.

3.1 (15 балів) Який з лінійних економетричнихмоделей оцінювання ви б віддали перевагу:

А) (5 балів) з незалежними змінними area, cluster, bread, parking;

Б) (5 балів) з незалежними змінними area, bread, parking;

В) (5 балів) з незалежними змінними area, cluster?

Подайте результати відповідних економетричнихоціноклінійної багатофакторної регресіїу кожному з випадків.

3.2 (5 балів) Виберіть найкращу з вашої точки зору модель. Відповідь обґрунтуйте.

In [237]:

```
#  Прогружаємо дані
df1 = pd.read_csv('data/task2_1.csv', names = ['Sales', 'Area', 'Cluster', 'Bread', 'Parking'])
df2 = pd.read_csv('data/task2_2.csv')

display(HTML(f'<h3> Зразок вхідних даних</h3>'))

df = pd.concat([df2,df1], axis = 0)
for col in df.columns:
    df[col] = df[col].apply(excel2python)

display(df.head())
```

Зразок вхідних даних

	Sales	Area	Cluster	Bread	Parking
0	1040090.60	107.0	4.0	1.0	1.0
1	795406.94	109.0	3.0	0.0	1.0
2	752838.26	112.0	3.0	1.0	0.0
3	808349.39	130.0	3.0	1.0	0.0
4	737827.97	150.0	2.0	0.0	1.0

3.1 (15 балів) Який з лінійних економетричнихмоделей оцінювання ви б віддали перевагу:

А) (5 балів) з незалежними змінними area, cluster, bread, parking;

Б) (5 балів) з незалежними змінними area, bread, parking;

В) (5 балів) з незалежними змінними area, cluster?

Подайте результати відповідних економетричних оцінок лінійної багатофакторної регресіїу кожному з випадків.

In [250]:

```
# A) (5 балів) з незалежними змінними area, cluster, bread, parking;

features = ['Area', 'Cluster', 'Bread', 'Parking']
X = sm.add_constant(df[features].values)
y = df.Sales.values.astype(float)

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Sales = ' + str(round(results.params[0]))+ ' + ' + ' + '.join([str(round(x))+'*'+f for x, f in zip(results.params[1:], features)]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.961
Model:                            OLS   Adj. R-squared:                  0.958
Method:                 Least Squares   F-statistic:                     327.6
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           1.12e-36
Time:                        21:12:49   Log-Likelihood:                -744.76
No. Observations:                  58   AIC:                             1500.
Df Residuals:                      53   BIC:                             1510.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -3.515e+05   4.78e+04     -7.348      0.000   -4.47e+05   -2.56e+05
x1          4181.0105    120.687     34.643      0.000    3938.942    4423.079
x2          2.483e+05   1.67e+04     14.835      0.000    2.15e+05    2.82e+05
x3         -3746.4380   3.99e+04     -0.094      0.926   -8.39e+04    7.64e+04
x4          5110.4959   4.21e+04      0.121      0.904   -7.94e+04    8.96e+04
==============================================================================
Omnibus:                        1.762   Durbin-Watson:                   2.160
Prob(Omnibus):                  0.414   Jarque-Bera (JB):                1.710
Skew:                          -0.393   Prob(JB):                        0.425
Kurtosis:                       2.702   Cond. No.                     1.25e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.25e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Відповідь:

```
Model:  Sales = -351473.0 + 4181.0*Area + 248309.0*Cluster + -3746.0*Bread + 5110.0*Parking
```

У даному випадку маємо модель, яка має досить непогану прогностичну силу. Можемо бачити, що коефіцієнт $R^2 = 0.958$, що є дуже добре.

Водночас може бачити також, що модель є надлишковою (деякі параметри є не значущими). Згідно з значеннями t статистики третій і четвертий параметри є не значущі для моделі. Бачимо що значення статистики є дуже мале, а також нуль входить в довірчий інтервал для цих параметрів.

In [251]:

```
# Б) (5 балів) з незалежними змінними area, bread, parking;

features = ['Area', 'Bread', 'Parking']
X = sm.add_constant(df[features].values)
y = df.Sales.values.astype(float)

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Sales = ' + str(round(results.params[0]))+ ' + ' + ' + '.join([str(round(x))+'*'+f for x, f in zip(results.params[1:], features)]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.800
Model:                            OLS   Adj. R-squared:                  0.789
Method:                 Least Squares   F-statistic:                     71.87
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           7.44e-19
Time:                        21:13:46   Log-Likelihood:                -792.30
No. Observations:                  58   AIC:                             1593.
Df Residuals:                      54   BIC:                             1601.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.05e+04   9.67e+04     -0.419      0.677   -2.34e+05    1.53e+05
x1          3834.0292    266.251     14.400      0.000    3300.228    4367.831
x2          3.981e+05    6.6e+04      6.028      0.000    2.66e+05     5.3e+05
x3          4.411e+05   6.79e+04      6.498      0.000    3.05e+05    5.77e+05
==============================================================================
Omnibus:                        3.525   Durbin-Watson:                   2.133
Prob(Omnibus):                  0.172   Jarque-Bera (JB):                1.784
Skew:                          -0.058   Prob(JB):                        0.410
Kurtosis:                       2.149   Cond. No.                     1.17e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.17e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Відповідь:

```
Model:  Sales = -40498.0 + 3834.0*Area + 398055.0*Bread + 441125.0*Parking
```

У даному випадку маємо модель, яка має гіршу прогностичну силу ніж перша модель. Можемо бачити, що коефіцієнт $R^2 = 0.789$, що є допустимо.

Водночас можемо бачити, що t статистика показує, що константна змінна є не значуща, так як значення статистики є дуже малим, що є не зовсім природня поведінка для такої моделі. Також довірчий інтервал для цієї константи є дуже великий при тому, що туди входить нуль. Такій моделі я б не довіряв, її ми відкидаємо.

In [252]:

```
# B) (5 балів) з незалежними змінними area, cluster?
features = ['Area', 'Cluster']
X = sm.add_constant(df[features].values)
y = df.Sales.values.astype(float)

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Sales = ' + str(round(results.params[0]))+ ' + ' + ' + '.join([str(round(x))+'*'+f for x, f in zi
p(results.params[1:], features)]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.961
Model:                            OLS   Adj. R-squared:                  0.960
Method:                 Least Squares   F-statistic:                     679.4
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           1.67e-39
Time:                        21:15:16   Log-Likelihood:                -744.78
No. Observations:                  58   AIC:                             1496.
Df Residuals:                      55   BIC:                             1502.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -3.525e+05   4.67e+04     -7.547      0.000   -4.46e+05   -2.59e+05
x1          4182.8436    117.134     35.710      0.000    3948.102    4417.585
x2          2.486e+05   9423.205     26.381      0.000     2.3e+05    2.67e+05
==============================================================================
Omnibus:                        1.812   Durbin-Watson:                   2.177
Prob(Omnibus):                  0.404   Jarque-Bera (JB):                1.758
Skew:                          -0.399   Prob(JB):                        0.415
Kurtosis:                       2.701   Cond. No.                     1.17e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.17e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Відповідь:

```
Model:  Sales = -352482.0 + 4183.0*Area + 248593.0*Cluster
```

У даному випадку маємо модель, яка має найкращу прогностичну силу порівнюючи дані нам три моделі. Можемо бачити, що коефіцієнт $R^2 = 0.960$, що є дуже добре.

Водночас можемо бачити, що t статистика показує, що всі змінні в моделі є значущі, так як значення t статистики є досить високим. Також довірчі інтервали для кожної з змінних є достатньо не великі та не містять нуля.

In [254]:

```
display(HTML(f'<h3>Відповідь: в результаті я б надав перевагу моделі  (B) Model:  Sales = -352482.0 + 4183.0*Area
+ 248593.0*Cluster </h3>'))
```

Відповідь: в результаті я б надав перевагу моделі (B) Model: Sales = -352482.0 + 4183.0*Area + 248593.0*Cluster

3.2 (5 балів) Виберіть найкращу з вашої точки зору модель. Відповідь обґрунтуйте.

Опісля проведеного аналізу найкраща, серед запропоновани, х на мій погляд є модель Sales = -352482.0 + 4183.0*Area* + *248593.0*Cluster. У цьому випадку я відкинув половину змінних, що були лишніми. В той час, ми маємо хороший коефіцієнт детермінації $R^2 = 0.960$. Така модель показує хорошу точність, її можна дуже легко інтерпретувати: один метр квадратний підвищує продажі на 4183, вплив місця 248593.

Можна також було б спробувати і нелінійні моделі (дерево рішень, випадковий ліс). Проте у них є більша ймовірність перевчитися під дані, яких у нас до того ж не багато.

Також спробував лінійну регресивну модель з відсутнім вільним членом для різних комбінацій з параметрів. В результаті використав би напевно її (також тільки з параметрими 'Area', 'Cluster'). Вона має кращий показник R^2 та навіть ще краще інтерпритується

In [274]:

```
features = ['Area', 'Cluster']
X = df[features].values
y = df.Sales.values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Sales = ' + ' + '.join([str(round(x))+'*'+f for x, f in zip(results.params, features)]))
```

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.991
Model:                            OLS   Adj. R-squared (uncentered):              0.991
Method:                 Least Squares   F-statistic:                              3113.
Date:                Wed, 07 Oct 2020   Prob (F-statistic):                    4.02e-58
Time:                        22:09:42   Log-Likelihood:                         -765.39
No. Observations:                  58   AIC:                                      1535.
Df Residuals:                      56   BIC:                                      1539.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          3382.4885     70.313     48.106      0.000    3241.634    3523.343
x2          1.949e+05   8736.101     22.310      0.000    1.77e+05    2.12e+05
==============================================================================
Omnibus:                        0.746   Durbin-Watson:                   1.694
Prob(Omnibus):                  0.689   Jarque-Bera (JB):                0.764
Skew:                          -0.000   Prob(JB):                        0.683
Kurtosis:                       2.438   Cond. No.                         154.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

```
Model:  Sales = 3382.0*Area + 194899.0*Cluster
```

Завдання 4

Компанія звиробництва будівельних сумішей розглядає можливість продати бізнес стратегічному інвестору. Інвестиційна консалтингова компанія для оцінки вартості бізнесу розглянула аналогічний бізнес у подібних країнах і його характеристики:

1)вартість активів;

2)чисті річні продажі;

3)річний показник EBITDA.

В результаті консультанти отримали такий масив даних.

Як би ви оцінили вартість активів цієї компанії, якщо відомо, що її NETSALES=150 млн. доларів, а EBITDA=8.8млн. доларів?

Використайте лінійну регресивну модель з відсутнім вільним членом (константа).

4.1 Дослідіть три варіанти моделі:

1) залежна змінна –eva, незалежна змінна –netsales;

2) залежна змінна –eva, незалежна змінна –ebitda;

3) залежна змінна –eva, дві незалежні змінні –netsales, ebitda.

4.2 (5 балів) Яка з моделей,на ваш погляд,є найкращою?

In [323]:

```
display(HTML(f'<h3> Вхідні дані</h3>'))
df = pd.read_csv('data/task4.csv')
df = df[['EVA', 'NET Sales', 'EBITDA']]
df
```

Вхідні дані

Out[323]:

	EVA	NET Sales	EBITDA
0	18	60	2.5
1	15	43	2.1
2	21	75	4.1
3	65	140	11.0
4	80	200	15.0
5	70	180	10.0
6	130	390	25.0
7	45	120	7.5
8	60	1750	10.0
9	15	6	2.3
10	24	94	3.7
11	20	74	3.1
12	90	250	16.0
13	115	320	18.0
14	35	114	5.1
15	10	34	1.5
16	21	72	2.9
17	25	86	4.3
18	36	137	4.8
19	44	143	5.3
20	63	178	11.0
21	28	91	4.4
22	38	139	4.7
23	49	156	5.9
24	61	168	9.9

In [325]:

```
# 1) залежна змінна -eva, незалежна змінна -netsales;

features = ['NET Sales']
X = df[features].values
y = df.EVA.values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Sales = ' + ' + '.join([str(round(x,3))+'*'+f for x, f in zip(results.params, features)]))
print(f'Оцінка вартості активів цієї компанії за даною моделлю : {round(results.predict([[150]])[0],3)} млн. дола
рів')
```

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.352
Model:                            OLS   Adj. R-squared (uncentered):              0.325
Method:                 Least Squares   F-statistic:                              13.04
Date:                Sun, 11 Oct 2020   Prob (F-statistic):                     0.00140
Time:                        16:15:09   Log-Likelihood:                         -130.86
No. Observations:                  25   AIC:                                      263.7
Df Residuals:                      24   BIC:                                      264.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0871      0.024      3.612      0.001       0.037       0.137
==============================================================================
Omnibus:                       16.640   Durbin-Watson:                   0.848
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.099
Skew:                          -1.227   Prob(JB):                     2.15e-06
Kurtosis:                       7.363   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

```
Model:  Sales = 0.087*NET Sales
Оцінка вартості активів цієї компанії за даною моделлю : 13.067 млн. доларів
```

In [326]:

```
# 2) залежна змінна -eva, незалежна змінна -ebitda;

features = ['EBITDA']
X = df[features].values
y = df.EVA.values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Sales = ' + ' + '.join([str(round(x,3))+'*'+f for x, f in zip(results.params, features)]))
print(f'Оцінка вартості активів цієї компанії за даною моделлю : {round(results.predict([[8.8]])[0],3)} млн. дола
рів')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.985
Model:                            OLS   Adj. R-squared (uncentered):              0.984
Method:                 Least Squares   F-statistic:                              1544.
Date:                Sun, 11 Oct 2020   Prob (F-statistic):                    2.69e-23
Time:                        16:15:31   Log-Likelihood:                         -84.041
No. Observations:                  25   AIC:                                      170.1
Df Residuals:                      24   BIC:                                      171.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             5.8794      0.150     39.290      0.000       5.571       6.188
==============================================================================
Omnibus:                        5.644   Durbin-Watson:                   1.979
Prob(Omnibus):                  0.059   Jarque-Bera (JB):                3.794
Skew:                          -0.649   Prob(JB):                        0.150
Kurtosis:                       4.400   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

```
Model:  Sales = 5.879*EBITDA
Оцінка вартості активів цієї компанії за даною моделлю : 51.739 млн. доларів
```

In [327]:

```
# 3) залежна змінна -eva, дві незалежні змінні -netsales, ebitda.

features = ['NET Sales', 'EBITDA']
X = df[features].values
y = df.EVA.values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Sales = ' + ' + '.join([str(round(x,3))+'*'+f for x, f in zip(results.params, features)]))
print(f'Оцінка вартості активів цієї компанії за даною моделлю : {round(results.predict([[150, 8.8]])[0],3)} млн. доларів')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.985
Model:                            OLS   Adj. R-squared (uncentered):              0.984
Method:                 Least Squares   F-statistic:                              749.6
Date:                Sun, 11 Oct 2020   Prob (F-statistic):                    1.15e-21
Time:                        16:15:46   Log-Likelihood:                         -83.878
No. Observations:                  25   AIC:                                      171.8
Df Residuals:                      23   BIC:                                      174.2
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0026      0.005      0.550      0.588      -0.007       0.012
x2             5.8190      0.188     31.034      0.000       5.431       6.207
==============================================================================
Omnibus:                        4.270   Durbin-Watson:                   1.978
Prob(Omnibus):                  0.118   Jarque-Bera (JB):                2.489
Skew:                          -0.539   Prob(JB):                        0.288
Kurtosis:                       4.108   Cond. No.                         49.8
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

```
Model:  Sales = 0.003*NET Sales + 5.819*EBITDA
Оцінка вартості активів цієї компанії за даною моделлю : 51.59 млн. доларів
```

4.2 (5 балів) Яка з моделей,на ваш погляд,є найкращою?

В даному випадку я б обрав другу модель: Sales = 5.879*EBITDA

Дана модель має найкраще значення $R^2 = 0.984$. Таке ж значення має і модель 3, проте вона також має як видно не значущу змінну, так як для параметру NET Sales t статистика є дуже малою та довірчий інтервал для значення цього параметру містить нуль. Також для такої конфігурації маємо більший довірчий інтервал для EBITDA.

Отже, обираю модель Sales = 5.879*EBITDA, так як вона має найкраще значення $R^2 = 0.984$, не містить не значущих змінних та легко інтерпритується.

Використовуючи дану модель маємо оцінку компанії в 51.739 млн. доларів.

Завдання 5.

Керівництво фармацевтичної компанії Miller Inc. проводить дослідження того, як витрати на продажі та ціна на їх продукт впливають на обсяги продаж компанії в мільйонах упаковок цього продукту. Менеджмент компанії зібрав наступні дані:

Продажі, в млн. упаковок (Sales)	Витрати на продажі, в млн. $ (Exp)	Ціна, в $ (P)
6	2	10
4	1	11
16	8	12
10	5	13
12	6	14
8	4	15
12	7	16
16	9	17
14	8	18

Компанія хоче знайти вплив факторів витрат та ціни на обсяги продаж продукту в упаковках з допомогою лінійної регресивної моделі:

$$Sales = A + B \times Expenses + C \times Price + \epsilon$$

5.1 (5 балів) Знайдіть коефіцієнти A, B, та C в припущенні методу найменших квадратів (МНК).

5.2 (5 балів) Поясніть отримані результати для керівництва компанії Miller Inc.

In [292]:

```
sales = [6,4,16,10,12,8,12,16,14]
expences = [2,1,8,5,6,4,7,9,8]
price = [10,11,12,13,14,15,16,17,18]
df = pd.DataFrame({'sales':sales, 'expences':expences, 'price':price})
display(HTML(f'<h3> Вхідні дані</h3>'))
df
```

Вхідні дані

Out[292]:

	sales	expences	price
0	6	2	10
1	4	1	11
2	16	8	12
3	10	5	13
4	12	6	14
5	8	4	15
6	12	7	16
7	16	9	17
8	14	8	18

In [294]:

```
features = ['expences', 'price']
X = sm.add_constant(df[features].values)
y = df.sales.values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Sales = ' + str(round(results.params[0], 3))+ ' + ' + ' + '.join([str(round(x, 3))+'*'+f for x, f
in zip(results.params[1:], features)]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.994
Model:                            OLS   Adj. R-squared:                  0.992
Method:                 Least Squares   F-statistic:                     525.7
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           1.83e-07
Time:                        22:43:33   Log-Likelihood:                -2.0015
No. Observations:                   9   AIC:                             10.00
Df Residuals:                       6   BIC:                             10.59
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.0481      0.762      7.938      0.000       4.184       7.912
x1             1.7580      0.069     25.343      0.000       1.588       1.928
x2            -0.3519      0.071     -4.981      0.002      -0.525      -0.179
==============================================================================
Omnibus:                        8.385   Durbin-Watson:                   2.536
Prob(Omnibus):                  0.015   Jarque-Bera (JB):                3.171
Skew:                          -1.374   Prob(JB):                        0.205
Kurtosis:                       3.949   Cond. No.                         95.7
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

```
Model:  Sales = 6.048 + 1.758*expences + -0.352*price
```

Відповідь:

5.1 (5 балів) Знайдіть коефіцієнти A, B, та C в припущенні методу найменших квадратів (МНК).

A = 6.048, B = 1.758 та C = -0.352

5.2 (5 балів) Поясніть отримані результати для керівництва компанії Miller Inc.

Запропонована модель досить добре описує продажі компанії. Як ми можемо бачити додатковий 1 млн доларів витрат на продажі буде підвищувати продажі на 1.758 млн упаковок. В той ж час підвищення ціни на 1 доларів буде знижувати продажі 352000 одиниць товару.

Завдання 6

Бізнес діагностичної компанії побудований на задоволенні потреб клієнтів, які проходять медичні обстеження чи лікування і при цьому потребують додатково здати аналізи крові для уточнення діагнозу чи контролю ходу лікування. Також частину цільового сегменту складають клієнти, що у добровільному порядку хочуть зробити діагностику для профілактики різних типів захворювань. Роль територіального розміщення – ближче до основних потоків клієнтів та у зручних для них місцях – є одним з ключових факторів успіху бізнесу з забору біоматеріалів і подальшого їх аналізу. Кількість аналізів (замовлень), що постачає кожен маніпуляційний кабінет (МК) мережі залежить від кількох факторів:

1)Місце розташування МК: місто (x_1=1), районний центр (x_1=0);

2)Наявність окремого входу:власний вхід (x_2=1), приміщення в іншому медичному закладі (x_2=0);

3)Близькість до державної поліклініки: далеко (x_3=0), недалеко (x_3=1), близько (x_3=2);

4)Наявність поблизу МК конкурента: так (x_4=1), ні (x_4=0);

5)Якість персоналу з забору біоматеріалів (медсестри): поганий забір та некомунікабельність (x_5=0), добрий забір та некомунікабельність (x_5=1), добрий забір та комунікабельність (x_5=2);

6)Наявність потоку людей, що йдуть повз МК: невеликий потік (x_6=0), середній потік (x_6=1), великий потік (x_6=2);

7)Наявність рекламного біл-борду поблизу МК: немає -(x_7=0), є -(x_7=1).

Для мережі МК відомий набір значень факторів для кожного МК, місячні обсяги продаж та кількість пацієнтів:

№ МК	Кількість замовлень за місяць	фактор 1	фактор 2	фактор 3	фактор 4	фактор 5	фактор 6	фактор 7
1	1255	1	1	2	1	2	2	1
2	1180	1	1	2	1	2	2	1
3	1235	1	1	2	1	2	2	1
4	1075	1	0	1	1	2	2	1
5	945	1	0	1	1	2	1	1
6	755	1	0	2	0	2	1	0
7	525	1	0	0	0	1	0	0
8	687	1	0	0	1	1	2	0
9	1085	0	1	2	1	1	2	1
10	840	0	1	2	1	1	1	1
11	880	0	1	2	1	1	1	1
12	795	0	1	2	1	1	1	0
13	1160	0	1	2	1	2	2	1
14	925	0	1	2	1	1	2	1
15	890	0	1	2	1	1	1	1
16	850	0	1	2	1	1	1	1
17	800	0	0	1	1	2	1	0
18	655	0	0	1	1	1	1	0
19	485	0	0	1	0	1	1	0
20	645	0	0	1	0	1	1	0
21	515	0	0	1	1	1	1	0
22	730	1	0	1	0	2	1	0
23	580	0	0	1	0	1	1	0

6.1 (5 балів) Запишіть результати економетричного оцінювання впливу факторів 1-7 на кількість пацієнтів.

6.2 (10 балів) Якій з економетричних моделей оцінювання ви б віддали перевагу:

А) (2 бали) з незалежними змінними: фактор1, фактор 2, фактор 3, фактор 4, фактор 5, фактор 6, фактор 7;

Б) (2 бали) з незалежними змінними: фактор 1, фактор 2, фактор 3, фактор 4, фактор 5, фактор 6;

В) (2 бали) з незалежними змінними: фактор 1, фактор 3, фактор 4, фактор 5, фактор 6.

(4 бали) Відповідь обґрунтуйте на основі t-статистик та adjusted R-squared.

In [306]:

```
display(HTML(f'<h3> Вхідні дані</h3>'))
df = pd.read_csv('data/task6.csv',skiprows = 1, names = ['# МК', 'Кількість_замовлень', 'фактор1', 'фактор2', 'фактор3', 'фактор4', 'фактор5', 'фактор6', 'фактор7'])
df
```

Вхідні дані

Out[306]:

	# МК	Кількість_замовлень	фактор1	фактор2	фактор3	фактор4	фактор5	фактор6	фактор7
0	1	1255	1	1	2	1	2	2	1
1	2	1180	1	1	2	1	2	2	1
2	3	1235	1	1	2	1	2	2	1
3	4	1075	1	0	1	1	2	2	1
4	5	945	1	0	1	1	2	1	1
5	6	755	1	0	2	0	2	1	0
6	7	525	1	0	0	0	1	0	0
7	8	687	1	0	0	1	1	2	0
8	9	1085	0	1	2	1	1	2	1
9	10	840	0	1	2	1	1	1	1
10	11	880	0	1	2	1	1	1	1
11	12	795	0	1	2	1	1	1	0
12	13	1160	0	1	2	1	2	2	1
13	14	925	0	1	2	1	1	2	1
14	15	890	0	1	2	1	1	1	1
15	16	850	0	1	2	1	1	1	1
16	17	800	0	0	1	1	2	1	0
17	18	655	0	0	1	1	1	1	0
18	19	485	0	0	1	0	1	1	0
19	20	645	0	0	1	0	1	1	0
20	21	515	0	0	1	1	1	1	0
21	22	730	1	0	1	0	2	1	0
22	23	580	0	0	1	0	1	1	0

In [308]:

```
features = ['фактор1', 'фактор2', 'фактор3', 'фактор4', 'фактор5', 'фактор6', 'фактор7']
X = sm.add_constant(df[features].values)
y = df['Кількість_замовлень'].values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Кількість замовлень = ' + str(round(results.params[0], 3))+ ' + ' + ' + '.join([str(round(x, 3))+
'*'+f for x, f in zip(results.params[1:], features)]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.956
Model:                            OLS   Adj. R-squared:                  0.936
Method:                 Least Squares   F-statistic:                     46.98
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           4.43e-09
Time:                        23:08:38   Log-Likelihood:                -121.53
No. Observations:                  23   AIC:                             259.1
Df Residuals:                      15   BIC:                             268.1
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        271.9385     51.654      5.265      0.000     161.840     382.037
x1            24.4853     41.210      0.594      0.561     -63.352     112.323
x2           181.4262     64.367      2.819      0.013      44.232     318.621
x3           -13.5392     45.226     -0.299      0.769    -109.937      82.858
x4            29.6766     41.020      0.723      0.481     -57.755     117.108
x5           197.1302     43.621      4.519      0.000     104.154     290.107
x6           103.8212     29.868      3.476      0.003      40.159     167.484
x7           128.4037     43.900      2.925      0.010      34.833     221.975
==============================================================================
Omnibus:                        0.878   Durbin-Watson:                   2.938
Prob(Omnibus):                  0.645   Jarque-Bera (JB):                0.810
Skew:                           0.218   Prob(JB):                        0.667
Kurtosis:                       2.191   Cond. No.                         20.3
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

Model: Кількість замовлень = 271.938 + 24.485*фактор1 + 181.426*фактор2 + -13.539*фактор3 + 29.677*фактор4 + 197.13*фактор5 + 103.821*фактор6 + 128.404*фактор7

6.1 (5 балів) Запишіть результати економетричного оцінювання впливу факторів 1-7 на кількість пацієнтів.

Коментар: Даний аналіз показує, що з даними змінними можна побудувати модель з досить хорошим значенням R-squared = 0.956. Проте як ми можемо бачити, декілька факторів є не значущими, а саме фактори 1, 3 і 4.

6.2 (10 балів) Якій з економетричних моделей оцінювання ви б віддали перевагу:

А) (2 бали) з незалежними змінними: фактор1, фактор 2, фактор 3, фактор 4, фактор 5, фактор 6, фактор 7;

Б) (2 бали) з незалежними змінними: фактор 1, фактор 2, фактор 3, фактор 4, фактор 5, фактор 6;

В) (2 бали) з незалежними змінними: фактор 1, фактор 3, фактор 4, фактор 5, фактор 6.

(4 бали) Відповідь обґрунтуйте на основі t-статистик та adjusted R-squared.

In [309]:

```
# А) (2 бали) з незалежними змінними: фактор1, фактор 2, фактор 3, фактор 4, фактор 5, фактор 6, фактор 7;
features = ['фактор1', 'фактор2', 'фактор3', 'фактор4', 'фактор5', 'фактор6', 'фактор7']
X = sm.add_constant(df[features].values)
y = df['Кількість_замовлень'].values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Кількість замовлень = ' + str(round(results.params[0], 3))+ ' + ' + ' + '.join([str(round(x, 3))+
'*'+f for x, f in zip(results.params[1:], features)]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.956
Model:                            OLS   Adj. R-squared:                  0.936
Method:                 Least Squares   F-statistic:                     46.98
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           4.43e-09
Time:                        23:27:50   Log-Likelihood:                -121.53
No. Observations:                  23   AIC:                             259.1
Df Residuals:                      15   BIC:                             268.1
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        271.9385     51.654      5.265      0.000     161.840     382.037
x1            24.4853     41.210      0.594      0.561     -63.352     112.323
x2           181.4262     64.367      2.819      0.013      44.232     318.621
x3           -13.5392     45.226     -0.299      0.769    -109.937      82.858
x4            29.6766     41.020      0.723      0.481     -57.755     117.108
x5           197.1302     43.621      4.519      0.000     104.154     290.107
x6           103.8212     29.868      3.476      0.003      40.159     167.484
x7           128.4037     43.900      2.925      0.010      34.833     221.975
==============================================================================
Omnibus:                        0.878   Durbin-Watson:                   2.938
Prob(Omnibus):                  0.645   Jarque-Bera (JB):                0.810
Skew:                           0.218   Prob(JB):                        0.667
Kurtosis:                       2.191   Cond. No.                         20.3
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

Model: Кількість замовлень = 271.938 + 24.485*фактор1 + 181.426*фактор2 + -13.539*фактор3 + 29.677*фактор4 + 197.13*фактор5 + 103.821*фактор6 + 128.404*фактор7

In [310]:

```
# Б) (2 бали) з незалежними змінними: фактор 1, фактор 2, фактор 3, фактор 4, фактор 5, фактор 6;
features = ['фактор1', 'фактор2', 'фактор3', 'фактор4', 'фактор5', 'фактор6']
X = sm.add_constant(df[features].values)
y = df['Кількість_замовлень'].values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Кількість замовлень = ' + str(round(results.params[0], 3))+ ' + ' + ' + '.join([str(round(x, 3))+
'*'+f for x, f in zip(results.params[1:], features)]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.932
Model:                            OLS   Adj. R-squared:                  0.906
Method:                 Least Squares   F-statistic:                     36.26
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           1.92e-08
Time:                        23:28:15   Log-Likelihood:                -126.72
No. Observations:                  23   AIC:                             267.4
Df Residuals:                      16   BIC:                             275.4
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        226.1957     59.733      3.787      0.002      99.568     352.823
x1            45.8208     49.212      0.931      0.366     -58.505     150.146
x2           246.8847     73.226      3.372      0.004      91.652     402.117
x3            -2.2571     54.675     -0.041      0.968    -118.163     113.649
x4            67.9352     47.173      1.440      0.169     -32.066     167.937
x5           209.5294     52.677      3.978      0.001      97.859     321.200
x6           112.5341     36.060      3.121      0.007      36.091     188.977
==============================================================================
Omnibus:                        1.486   Durbin-Watson:                   2.723
Prob(Omnibus):                  0.476   Jarque-Bera (JB):                1.047
Skew:                           0.233   Prob(JB):                        0.593
Kurtosis:                       2.065   Cond. No.                         19.8
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

Model: Кількість замовлень = 226.196 + 45.821*фактор1 + 246.885*фактор2 + -2.257*фактор3 + 67.935*ф
актор4 + 209.529*фактор5 + 112.534*фактор6

In [311]:

```
# В) (2 бали) з незалежними змінними: фактор 1, фактор 3, фактор 4, фактор 5, фактор 6.
features = ['фактор1', 'фактор3', 'фактор4', 'фактор5', 'фактор6']
X = sm.add_constant(df[features].values)
y = df['Кількість_замовлень'].values

model = sm.OLS(y, X)
results = model.fit()
print(results.summary())

print()
print()
display(HTML(f'<h4>Відповідь:</h4>'))
print(f'Model:  Кількість замовлень = ' + str(round(results.params[0], 3))+ ' + ' + ' + '.join([str(round(x, 3))+
'*'+f for x, f in zip(results.params[1:], features)]))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.883
Model:                            OLS   Adj. R-squared:                  0.848
Method:                 Least Squares   F-statistic:                     25.62
Date:                Wed, 07 Oct 2020   Prob (F-statistic):           2.36e-07
Time:                        23:28:42   Log-Likelihood:                -132.89
No. Observations:                  23   AIC:                             277.8
Df Residuals:                      17   BIC:                             284.6
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        151.3074     70.354      2.151      0.046       2.874     299.741
x1           107.0276     58.037      1.844      0.083     -15.419     229.474
x2           154.5147     36.493      4.234      0.001      77.522     231.507
x3           140.2518     53.307      2.631      0.018      27.783     252.720
x4           113.6558     56.261      2.020      0.059      -5.045     232.357
x5           130.9476     45.225      2.896      0.010      35.532     226.363
==============================================================================
Omnibus:                        1.749   Durbin-Watson:                   3.107
Prob(Omnibus):                  0.417   Jarque-Bera (JB):                0.869
Skew:                          -0.470   Prob(JB):                        0.648
Kurtosis:                       3.153   Cond. No.                         12.7
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Відповідь:

```
Model:  Кількість замовлень = 151.307 + 107.028*фактор1 + 154.515*фактор3 + 140.252*фактор4 + 113.65
6*фактор5 + 130.948*фактор6
```

Коментар:

Третя модель є найгіршою по показнику $R^2 = 0.848$, також можемо бачити, що тільки фактор 3 і 6 в цій моделі є з достатньо значимим з значенням Т-статистики, для інших факторів ця статистика коливається навколо 2, що впринципі достатньо, проте не дуже впевнено.

Друга модель має краще значення $R^2 = 0.906$. Також бачимо, що вже 3 фактори і константа мають значення Т-статистики більше трьох, що означає їхню значемість для моделі.

Перша модель має $R^2 = 0.936$, що є найкращим значенням серед трьох моделей. Проте не набагато краще за другу модель. Можемо бачити, однаково з другою моделлю має 3 не значимих факторів, інші є значемі відповідно до Т-статистики. (близько до/або >3)

Відповідь: Відповідно до проведеного вище аналізу я б рекомендував обрати першу модель за рахунок найкращого значення R^2 та значень Т-статистики.

In [ ]:

```

```