Група 4:

Криволап Дмитро

Остапюк Наталія

Свердлов Денис

Шовковий Владислав

Завдання 2

У таблиці наведено дані ДержСтату про основні показники ведення лісогосподарської діяльності.

1. Провести дескриптивний та візуальний аналіз:

a. Обчислити вибіркові характеристики (середнє, медіана, дисперсія, інтерквартильний розмах), знайти вибіркову кореляційну матрицю.

b. Для наборів спостережень побудувати графіки типу «вусатих коробочок»

c. Побудувати матричну діаграму розсіювання, обрати залежну величину Y та набір факторів X1,X2,…,Xp для подальшого регресійного аналізу.

d. Оцінити параметри лінійної регресії залежності величини (Y) від обраних факторів (X) та проаналізувати результати (адекватність, можливість побудови прогнозу).

In [133]:
```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn
from copy import deepcopy

data = pd.read_csv('data.csv', encoding='ISO-8859-1', sep = ";")
data.columns = ['Роки','Обсяги продукції','Площа рубок','Заготівля деревини','Кількість пожеж','Площа спалених земель','З

data
```

Out[133]:

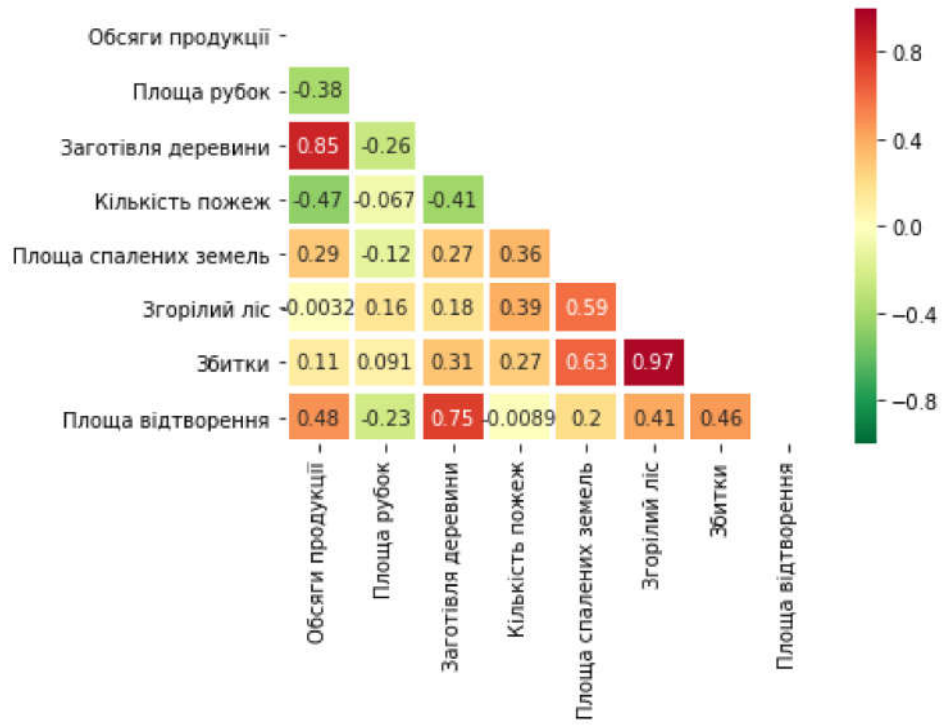
	Роки	Обсяги продукції	Площа рубок	Заготівля деревини	Кількість пожеж	Площа спалених земель	Згорілий ліс	Збитки	Площа відтворення
0	1997	373.0	403100	10597.0	2309	1467	11806	615.4	38.5
1	1998	396.6	435400	10548.7	3915	4418	123034	4555.7	36.7
2	1999	521.3	434600	10308.7	6070	5532	163858	5822.3	38.6
3	2000	744.4	455100	11261.7	3696	1610	20249	1367.6	37.8
4	2001	824.2	570300	12022.3	3205	3772	139604	6204.3	42.6
5	2002	946.8	376591	12826.8	6383	4983	59206	3378.9	45.9
6	2003	1108.9	383191	15953.3	4527	2817	19720	1817.5	48.3
7	2004	1594.6	468648	17300.7	1876	595	1944	428.7	53.9
8	2005	1991.1	484673	17124.3	4223	2325	32101	3535.0	58.6
9	2006	2451.1	468188	17759.8	3842	4287	53119	5917.6	66.7
10	2007	2956.3	476241	19013.9	6100	13787	1304271	188412.0	73.6
11	2008	3382.7	425344	17687.5	4042	5529	395257	58750.3	80.2
12	2009	3138.1	357949	15876.5	7036	6315	223764	24686.4	80.9
13	2010	4097.7	402205	18064.6	3240	3668	343840	26728.4	70.1
14	2011	5674.8	421750	19746.2	2526	1049	11804	3215.9	72.4
15	2012	5911.6	417005	19763.6	2163	3479	289291	56062.7	70.1
16	2013	6363.9	415420	20340.6	1113	418	2496	1376.2	67.7
17	2014	7739.9	382623	20672.4	2003	13778	144694	51701.8	58.0
18	2015	10778.2	399296	21924.2	3813	14691	170686	20164.5	60.4
19	2016	12838.8	386382	22612.8	1249	1249	32559	8619.2	63.2

а. Обчислити вибіркові характеристики (середнє, медіана, дисперсія, інтерквартильний розмах), знайти вибіркову кореляційну матрицю.

In [85]:
```
#Correlation heatmap
corr = data.iloc[:,1:].corr()
mask = np.zeros_like(corr)
mask[np.triu_indices_from(mask)] = True
seaborn.heatmap(corr, cmap = 'RdYlGn_r', vmax = 1., vmin = -1., mask = mask, linewidth = 2.5,annot=True)

summary = data.iloc[:,1:].describe()
summary.loc['Interquantile_range'] = (summary.iloc[6,:] - summary.iloc[4,:]).tolist()
data1 = deepcopy(data.iloc[:,1:])
summary
```

Out[85]:

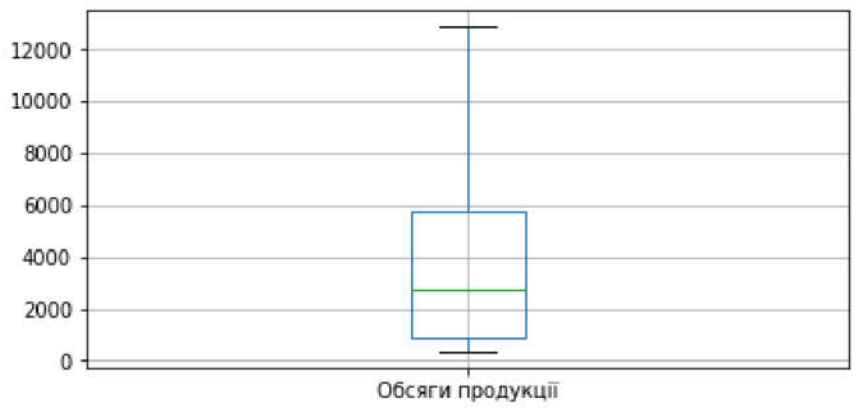
	Обсяги продукції	Площа рубок	Заготівля деревини	Кількість пожеж	Площа спалених земель	Згорілий ліс	Збитки	Площа відтворення
count	20.000000	20.000000	20.000000	20.00000	20.000000	2.000000e+01	20.000000	20.000000
mean	3691.700000	428200.300000	16570.280000	3666.55000	4788.450000	1.771651e+05	23668.020000	58.210000
std	3548.543466	49028.002296	3994.064643	1714.86395	4364.899715	2.904296e+05	43248.396435	14.646102
min	373.000000	357949.000000	10308.700000	1113.00000	418.000000	1.944000e+03	428.700000	36.700000
25%	916.150000	396067.500000	12625.675000	2272.50000	1574.250000	2.011675e+04	2866.300000	45.075000
50%	2703.700000	419377.500000	17494.100000	3754.50000	3720.000000	9.112000e+04	5869.950000	59.500000
75%	5734.000000	458372.000000	19750.550000	4299.00000	5529.750000	1.839555e+05	25196.900000	70.100000
max	12838.800000	570300.000000	22612.800000	7036.00000	14691.000000	1.304271e+06	188412.000000	80.900000
Interquantile_range	4817.850000	62304.500000	7124.875000	2026.50000	3955.500000	1.638388e+05	22330.600000	25.025000

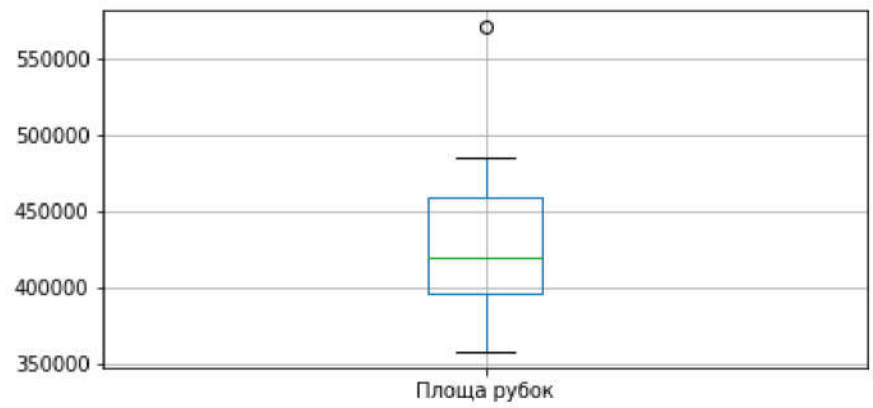


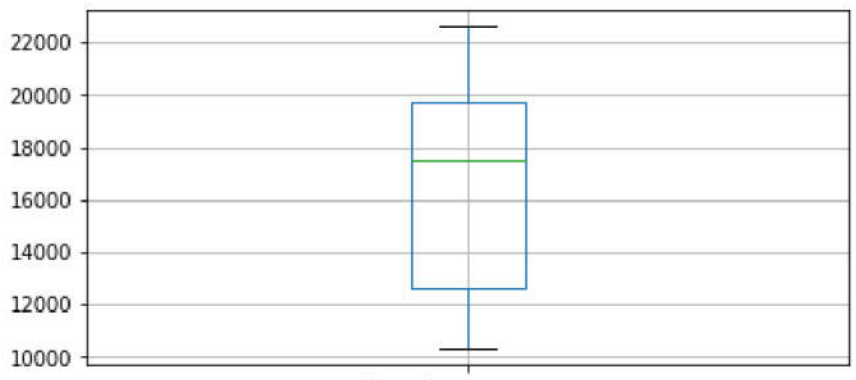
Видно дуже високу кореляцію між збитками і об'ємом згорілого лісу, а також між обсягом продукції і об'ємом заготовленої деревини.

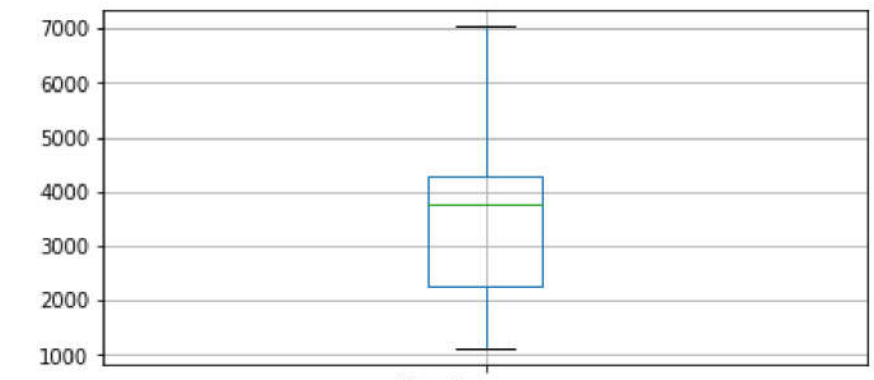
Вусаті коробочки

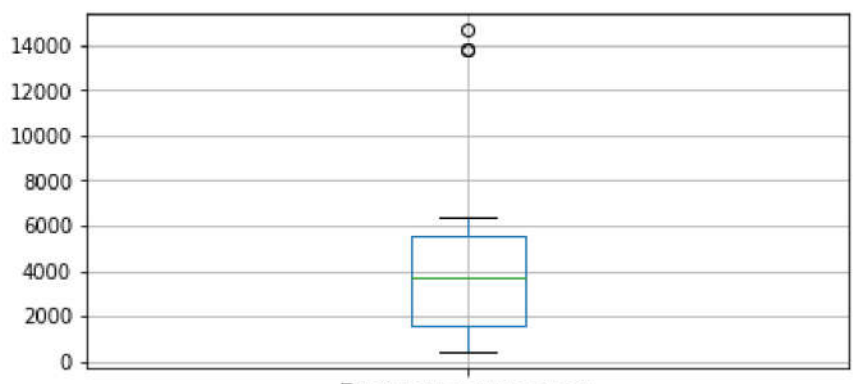
In [119]:
```
fig, axis = plt.subplots(4, 2, figsize=(15, 15))
i = 0
for col in data1.columns:
    data1.boxplot(col,ax = axis[int(i/ 2),i %2])
    i+=1
```

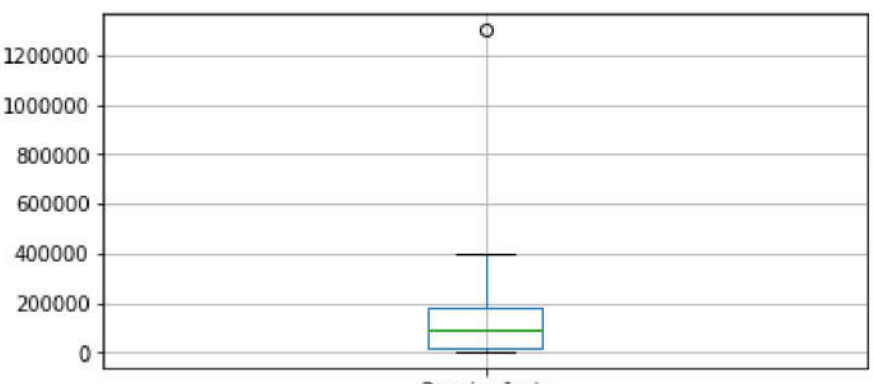


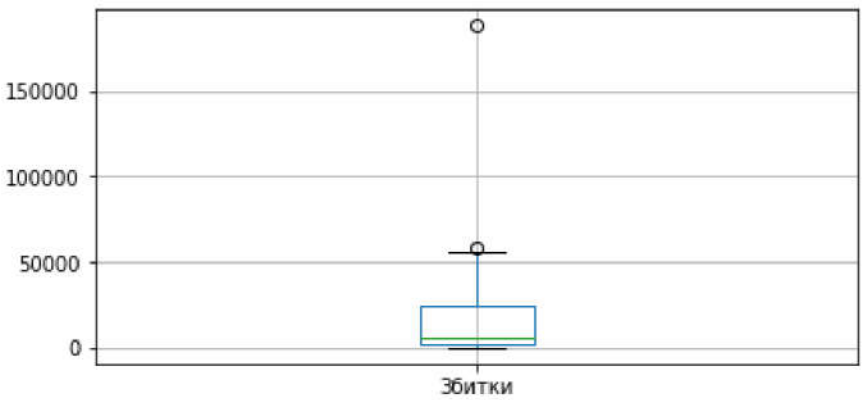


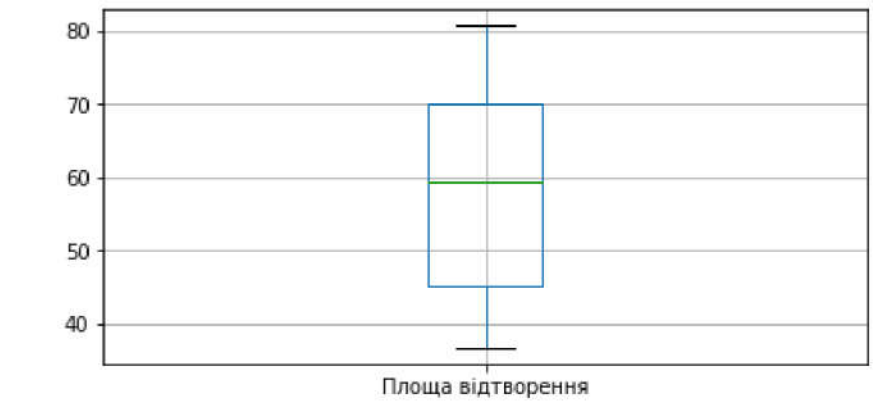






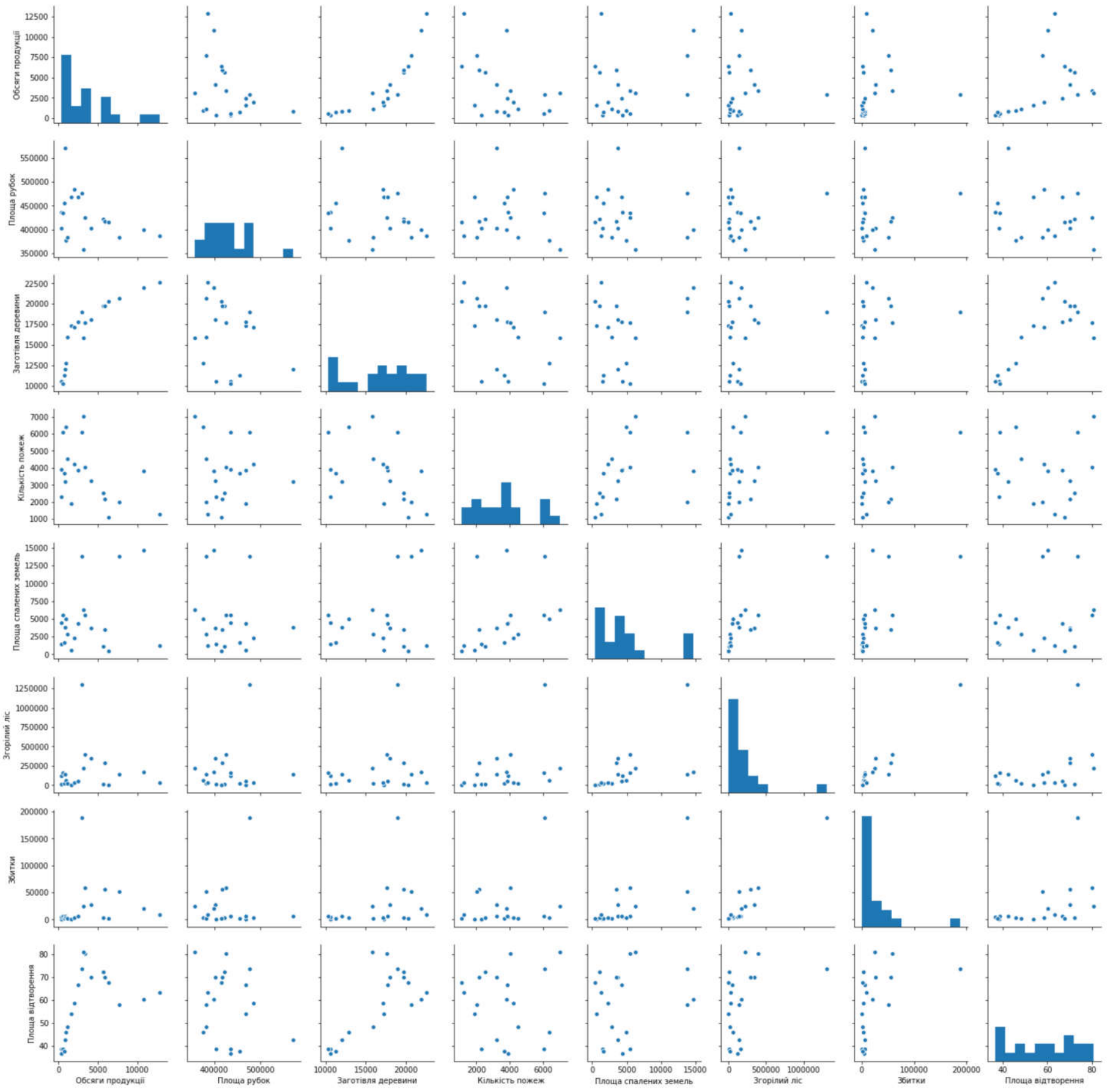






In [121]:
```
plt.figure(figsize=(11,6))
seaborn.pairplot(data.iloc[:,1:9])
```

Out[121]:
```
<seaborn.axisgrid.PairGrid at 0x1bf443569e8>

<matplotlib.figure.Figure at 0x1bf4372ecc0>
```

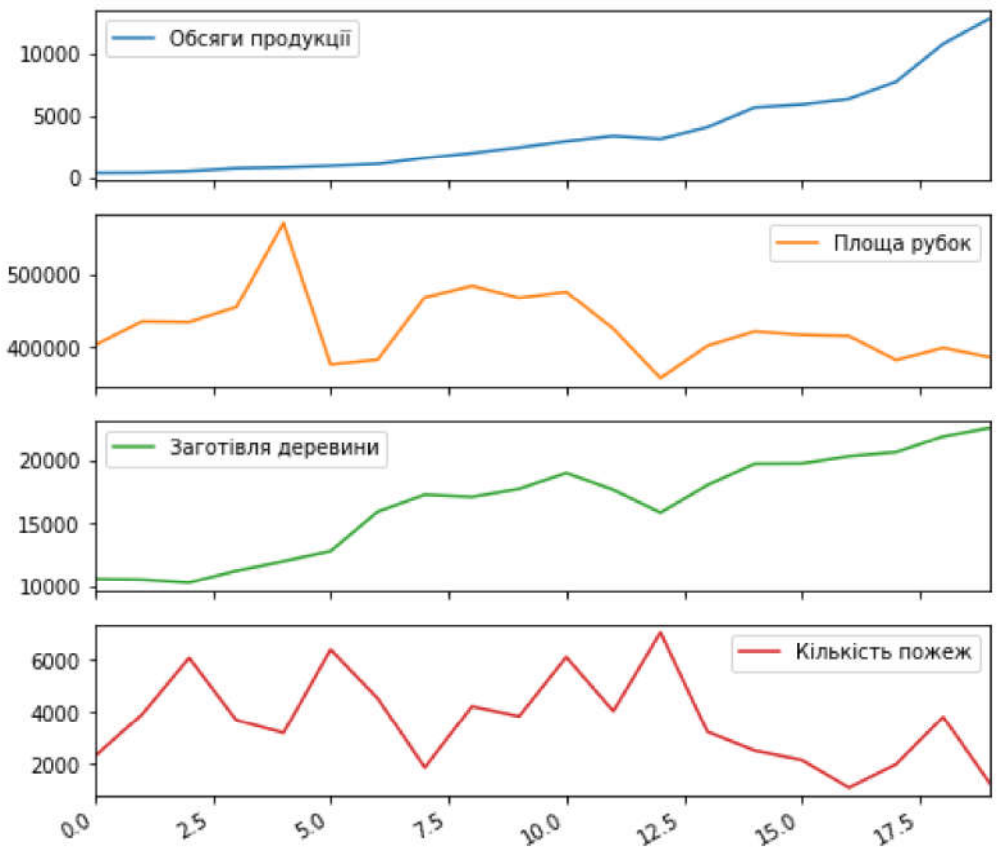


З діаграм розсіювання явно можна побачити тільки залежність обсягу продукції від об'єму заготівлі деревини.

Часові ряди

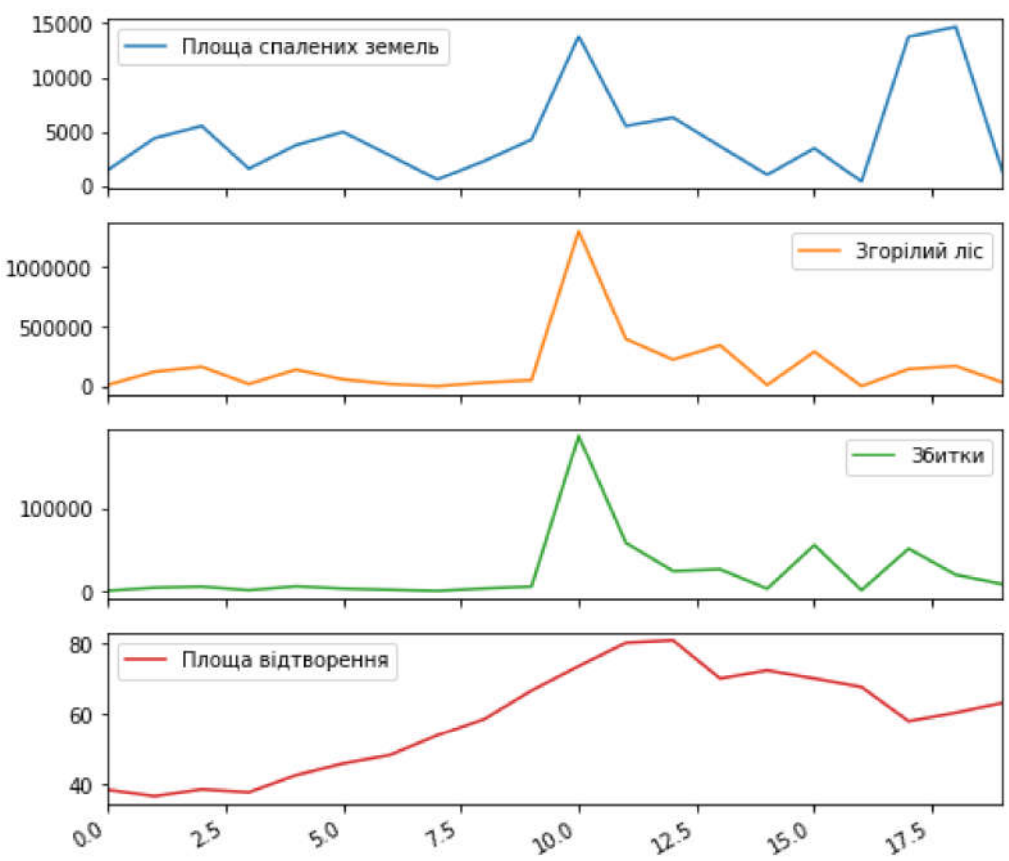
In [24]:
```
data.iloc[:,1:5].plot(subplots=True, figsize=(8, 8)); plt.legend(loc='best')
```

Out[24]: <matplotlib.legend.Legend at 0x1bf3db436a0>



In [25]:
```
data.iloc[:,5:].plot(subplots=True, figsize=(8, 8)); plt.legend(loc='best')
```

Out[25]: <matplotlib.legend.Legend at 0x1bf3da00cc0>



Вибір залежної змінної

Очевидно, що **обсяг продукції в гривнях** знаходиться в прямій залежності від **об'єму заготовленої деревини** - якщо ми знаємо об'єм заготовленої деревини, то нам уже не важливі дані про пожежі й інші фактори, які в нас наявні, обсяг продукції в гривнях буде приблизно дорівнювати об'єму заготівлі, помноженому на вартість кубометру деревини(в залежності від виду продукції), про яку даних у нас немає. Тому є сенс передбачати саме **об'єм заготівлі деревини**. Також ми робимо припущення, що на план об'єму заготівлі впливає обсяг реалізованої продукції в минулому році - тому додаємо цю змінну як один із можливих факторів.

Відбір факторів для регресії

Зразу відкидаємо фактор **'Збитки'**, оскільки він сильно корелює з фактором **'Згорілий ліс'**(коефіцієнт кореляції - 0.97). Додаємо як можливий фактор **'лаг'** - обсяг реалізованої продукції в минулому році

Відбір факторів для моделі проведемо з точки зору максимізації **R-квадрат** (коефіцієнт детермінації - частка дисперсії залежної змінної, яка пояснюється моделлю залежності) та мінімазації **AIC** (інформаційний критерій Акаіке - оцінює кількість втраченої інформації для даної моделі)

Оскільки кількість факторів невелика, то вибір моделі можна зробити за допомогою повного перебору всіх можливих комбінацій факторів.

В результаті були відібрані наступні фактори: **'Площа відтворення', 'Кількість пожеж', 'Площа спалених земель', 'лаг'**

Заготівля деревини

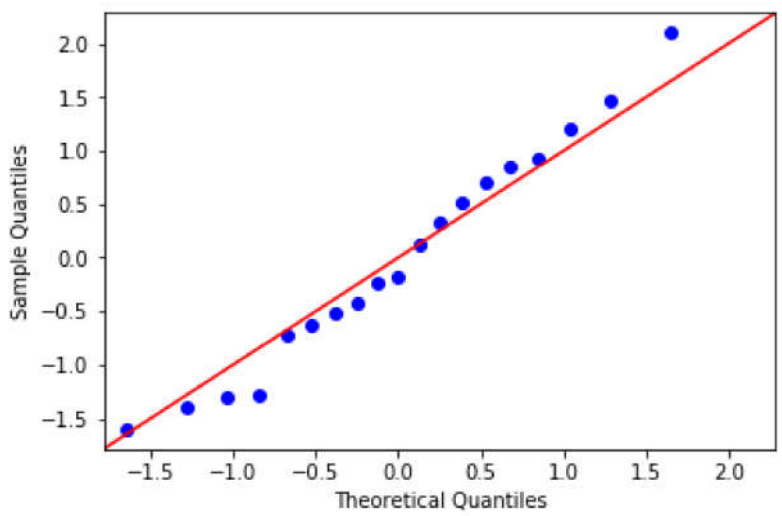
In [159]:
```
X = data[['Площа відтворення', 'Кількість пожеж','Площа спалених земель','lag']]
X = sm.add_constant(X)
model = sm.OLS(data['Заготівля деревини'], X, missing='drop')
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Заготівля деревини   R-squared:                       0.888
Model:                            OLS   Adj. R-squared:                  0.855
Method:                 Least Squares   F-statistic:                     27.61
Date:                Mon, 25 Jun 2018   Prob (F-statistic):           1.64e-06
Time:                        15:34:11   Log-Likelihood:                -162.51
No. Observations:                  19   AIC:                             335.0
Df Residuals:                      14   BIC:                             339.7
Df Model:                           4
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     9627.7567   1720.491      5.596      0.000    5937.671    1.33e+04
Площа відтворення          128.9499     28.591      4.510      0.000      67.628     190.271
Кількість пожеж             -0.7619      0.291     -2.617      0.020      -1.386      -0.138
Площа спалених земель        0.1676      0.096      1.745      0.103      -0.038       0.374
lag                          0.5087      0.187      2.713      0.017       0.107       0.911
==============================================================================
Omnibus:                        0.979   Durbin-Watson:                   0.924
Prob(Omnibus):                  0.613   Jarque-Bera (JB):                0.861
Skew:                           0.291   Prob(JB):                        0.650
Kurtosis:                       2.135   Cond. No.                     4.11e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.11e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

C:\Users\NO\Anaconda3\lib\site-packages\scipy\stats\stats.py:1390: UserWarning: kurtosistest only valid for n>=20 ... c
ontinuing anyway, n=19
  "anyway, n=%i" % int(n))
```

Перевірка умов теореми Гауса-Маркова

Нормальність залишків

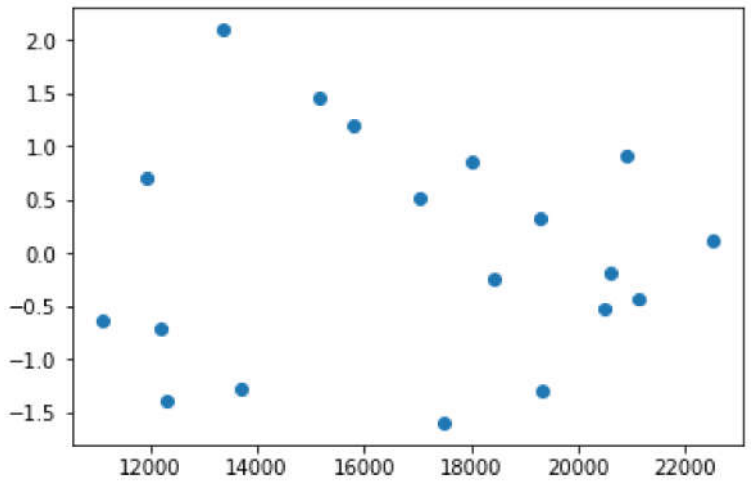
In [162]:
```
from statsmodels.stats.outliers_influence import OLSInfluence
from statsmodels.graphics.gofplots import qqplot
from statsmodels.stats.diagnostic import het_breuschpagan
from statsmodels.stats.diagnostic import het_white
resid_st_x1 = OLSInfluence(results).resid_studentized_external
fig = qqplot(resid_st_x1, line='45')
plt.show()
```



Однорідність залишків

In [34]:
```
plt.scatter(results.fittedvalues, resid_st_x1)
```

Out[34]:
```
<matplotlib.collections.PathCollection at 0x1bf408d1e80>
```



In [166]:
```
print("Тест Уайта lm_pvalue: ",het_white(results.resid, results.model.exog)[1])
```

```
Тест Уайта lm_pvalue:  0.45773077475083845
```

За результатами теста Уайта не відхиляємо гіпотезу про однорідність залишків

Автокореляція залишків

Статистика критерій Дарбіна-Уотсона дорівнює 0.924, що не дозволяє прийняти гіпотезу про некорельованість залишків

Висновки про адекватність моделі

Оскільки не виконується умова про незалежність залишків, то не можна стверджувати, що знайдена модель лінійної регресії адекватно оцінює залежну змінну. Можливо, це пояснюється тим, що об'єм заготовленої деревини залежить зовсім від інших факторів (наприклад від контрактів, які були укладені зі підприємствами-споживачами деревини), або характер залежності нелінійний.