Regr

March 31, 2022

[1]:
```
"Regresjons analyse av Skattetall Oslo 2020"
"Hentet ut fra vg"
```

[1]:
```
'Hentet ut fra vg'
```

[2]:
```
from bs4 import BeautifulSoup
import requests
import numpy as np
```

[3]:
```
def fetch_html_tables(url):
    "Returns a list of tables in the html of url"
    page = requests.get(url)
    bs=BeautifulSoup(page.content)
    tables=bs.find_all('table')
    return tables

tables=fetch_html_tables('https://www.vg.no/spesial/skattelister/2020/0301/')
table_html=tables[0]

#printing top
print(str(table_html)[:1000])
```

```
<table class="table table-sm"><thead><tr><th class="w-5"></th><th
class="w-30">Navn</th><th class="text-right clickable">Inntekt<i
class="material-icons md-14 middle"> </i></th><th class="text-right
clickable">Formue</th><th class="text-right
clickable">Skatt</th></tr></thead><tbody><tr><td>1<!-- -->.</td><td><div
class="name">ØYSTEIN STRAY<!-- --> <!-- -->SPETALEN</div><div class="text-
muted"><a class="text-muted" href="/spesial/skattelister/2020/0301/">Oslo</a>,
f.<!-- --> <!-- -->1962</div></td><td class="text-right">258 709 731</td><td
class="text-right">2 586 723 237</td><td class="text-
right">105 131 374</td></tr><tr><td>2<!-- -->.</td><td><div
class="name">ØYSTEIN<!-- --> <!-- -->MOAN</div><div class="text-muted"><a
class="text-muted" href="/spesial/skattelister/2020/0301/">Oslo</a>, f.<!--
--> <!-- -->1959</div></td><td class="text-right">181 404 420</td><td
class="text-right">961 391 825</td><td class="text-
right">52 228 271</td></tr><tr><td>3<!-- -->.</td><td><div class="name
```

[4]:
```
def html_to_table(html):
    "Returns the table defined in html as a list"
    #defining the table:
    table=[]
    #iterating over all rows
    for row in html.find_all('tr'):
        r=[]
        #finding all cells in each row:
        cells=row.find_all('td')

        #if no cells are found, look for headings
        if len(cells)==0:
            cells=row.find_all('th')

        #iterate over cells:
        for cell in cells:
            cell=format(cell)
            r.append(cell)

        #append the row to t:
        table.append(r)
    return table

def format(cell):
    "Returns a string after converting bs4 object cell to clean text"
    if cell.content is None:
        s=cell.text
    elif len(cell.content)==0:
        return ''
    else:
        s=' '.join([str(c) for c in cell.content])

    #here you can add additional characters/strings you want to
    #remove, change punctuations or format the string in other
    #ways:
    s=s.replace('\xa0',"")
    s=s.replace('\n',"")
    s=s.replace("\ue5cf","")
    return s

table=html_to_table(table_html)

#printing top
print(str(table)[:1000])
```

```
[['', 'Navn', 'Inntekt', 'Formue', 'Skatt'], ['1.', 'ØYSTEIN STRAY SPETALENOslo,
f.1962', '258709731', '2586723237', '105131374'], ['2.', 'ØYSTEIN MOANOslo,
f.1959', '181404420', '961391825', '52228271'], ['3.', 'MAGNUS REITANOslo,
f.1975', '166542768', '4620795695', '92706050'], ['4.', 'OLE ROBERT REITANOslo,
f.1971', '162833343', '4546957558', '90924296'], ['5.', 'MARGARET BOEL
GARMANNOslo, f.1955', '154930624', '2562173467', '70855059'], ['6.', 'JØRGEN
DAHLOslo, f.1969', '131447948', '2790325759', '65814357'], ['7.', 'EDGAR
HAUGENOslo, f.1965', '112271516', '2777922578', '58914860'], ['8.', 'ERIK WILSON
LANDGRAFFOslo, f.1984', '109748661', '100233911', '1459975'], ['9.', 'TOR ØIVIND
FJELDOslo, f.1979', '103320599', '2743633213', '55434237'], ['10.', 'EILERT
GIERTSEN HANOAOslo, f.1970', '101504287', '78400112', '33184365'], ['11.', 'ERIK
KRISTOFFER ARTHUROslo, f.1962', '90443514', '194726696', '30669028'], ['12.',
'ODD JOHNNY WINGEOslo, f.1975', '89197038', '441465476', '31968678'],
```

```
[5]: ';'.join(table[0])
```

```
[5]: ';Navn;Inntekt;Formue;Skatt'
```

```
[ ]:
```

```
[6]: def save_data(file_name,table):
         "Saves table to file_name"
         f=open(file_name,'w')
         for row in table:
             f.write(';'.join(row)+'\n')
         f.close()

     save_data('Skatteliste_oslo.csv',table)
```

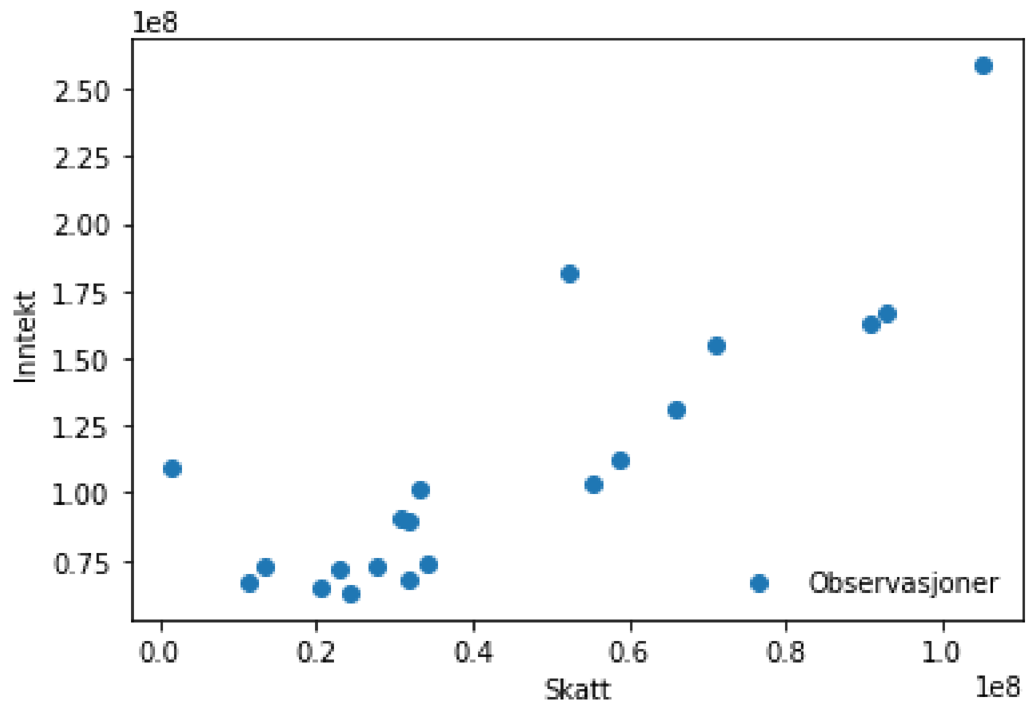
```
[7]: import pandas as pd
     oslo = pd.read_csv('Skatteliste_oslo.csv', delimiter=';', encoding='utf8')

     oslo
```

```
[7]:     Unnamed: 0                              Navn    Inntekt  \
     0          1.0      ØYSTEIN STRAY SPETALENOslo, f.1962  258709731
     1          2.0               ØYSTEIN MOANOslo, f.1959  181404420
     2          3.0              MAGNUS REITANOslo, f.1975  166542768
     3          4.0          OLE ROBERT REITANOslo, f.1971  162833343
     4          5.0        MARGARET BOEL GARMANNOslo, f.1955  154930624
     5          6.0                JØRGEN DAHLOslo, f.1969  131447948
     6          7.0               EDGAR HAUGENOslo, f.1965  112271516
     7          8.0        ERIK WILSON LANDGRAFFOslo, f.1984  109748661
     8          9.0             TOR ØIVIND FJELDOslo, f.1979  103320599
     9         10.0        EILERT GIERTSEN HANOAOslo, f.1970  101504287
     10        11.0      ERIK KRISTOFFER ARTHUROslo, f.1962   90443514
     11        12.0             ODD JOHNNY WINGEOslo, f.1975   89197038
     12        13.0               RAKESH PATELOslo, f.1975   73778983
     13        14.0              PÅL ERIK SJÅTILOslo, f.1972   72857445
```

```
14        15.0             STEIN ERIK HAGENOslo, f.1956   72780634
15        16.0               KENNETH LØVOLDOslo, f.1970   71412906
16        17.0  KRISTIAN GJERDRUM ROSENBERGOslo, f.1962   67976007
17        18.0   NICOLAI HARALD LØVENSKIOLDOslo, f.1962   66690101
18        19.0                 PETTER SOLUMOslo, f.1954   64567647
19        20.0             KARL JOHAN SUNDEOslo, f.1949   62920580

        Formue      Skatt
0   2586723237  105131374
1    961391825   52228271
2   4620795695   92706050
3   4546957558   90924296
4   2562173467   70855059
5   2790325759   65814357
6   2777922578   58914860
7    100233911    1459975
8   2743633213   55434237
9     78400112   33184365
10   194726696   30669028
11   441465476   31968678
12    13426780   34149789
13   188862505   13496248
14   433286845   27864669
15           0   23024867
16    53784556   31891947
17   103552557   11340325
18    28793539   20667701
19   542148645   24335298
```

[17]:
```python
import numpy as np
from matplotlib import pyplot as plt

fig,ax=plt.subplots()

#adding axis lables:
ax.set_ylabel('Inntekt')
ax.set_xlabel('Skatt')

#plotting the function:
ax.scatter(oslo['Skatt'], oslo['Inntekt'],  label='Observasjoner')
ax.legend(loc='lower right',frameon=False)
```

[17]:
```
<matplotlib.legend.Legend at 0x7fd22f041bb0>
```



[9]:
```
y=oslo['Skatt']
pd.DataFrame(y)
```

[9]:
```
        Skatt
0   105131374
1    52228271
2    92706050
3    90924296
4    70855059
5    65814357
6    58914860
7     1459975
8    55434237
9    33184365
10   30669028
11   31968678
12   34149789
13   13496248
14   27864669
15   23024867
16   31891947
17   11340325
18   20667701
19   24335298
```

```
[10]: x=pd.DataFrame((oslo['Skatt']))
      x['Intercept']=1
      x
```

```
[10]:        Skatt  Intercept
      0   105131374          1
      1    52228271          1
      2    92706050          1
      3    90924296          1
      4    70855059          1
      5    65814357          1
      6    58914860          1
      7     1459975          1
      8    55434237          1
      9    33184365          1
      10   30669028          1
      11   31968678          1
      12   34149789          1
      13   13496248          1
      14   27864669          1
      15   23024867          1
      16   31891947          1
      17   11340325          1
      18   20667701          1
      19   24335298          1
```

```
[11]: from statsmodels.regression.linear_model import OLS

      res=OLS(y,x).fit()

      print(res.summary())
```

```
/usr/local/Miniconda3-py39_4.10.3-Linux-x86_64/lib/python3.9/site-
packages/statsmodels/compat/pandas.py:65: FutureWarning: pandas.Int64Index is
deprecated and will be removed from pandas in a future version. Use pandas.Index
with the appropriate dtype instead.
  from pandas import Int64Index as NumericIndex

                            OLS Regression Results
==============================================================================
Dep. Variable:                  Skatt   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 2.052e+31
Date:                Tue, 29 Mar 2022   Prob (F-statistic):         5.70e-272
Time:                        10:59:33   Log-Likelihood:                 320.52
No. Observations:                  20   AIC:                            -637.0
Df Residuals:                      18   BIC:                            -635.1
Df Model:                           1
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Skatt          1.0000   2.21e-16   4.53e+15      0.000       1.000       1.000
Intercept   4.657e-09   1.15e-08      0.404      0.691   -1.95e-08    2.88e-08
==============================================================================
Omnibus:                        1.681   Durbin-Watson:                   0.173
Prob(Omnibus):                  0.431   Jarque-Bera (JB):                1.405
Skew:                          -0.515   Prob(JB):                        0.495
Kurtosis:                       2.209   Cond. No.                     9.61e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[2] The condition number is large, 9.61e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

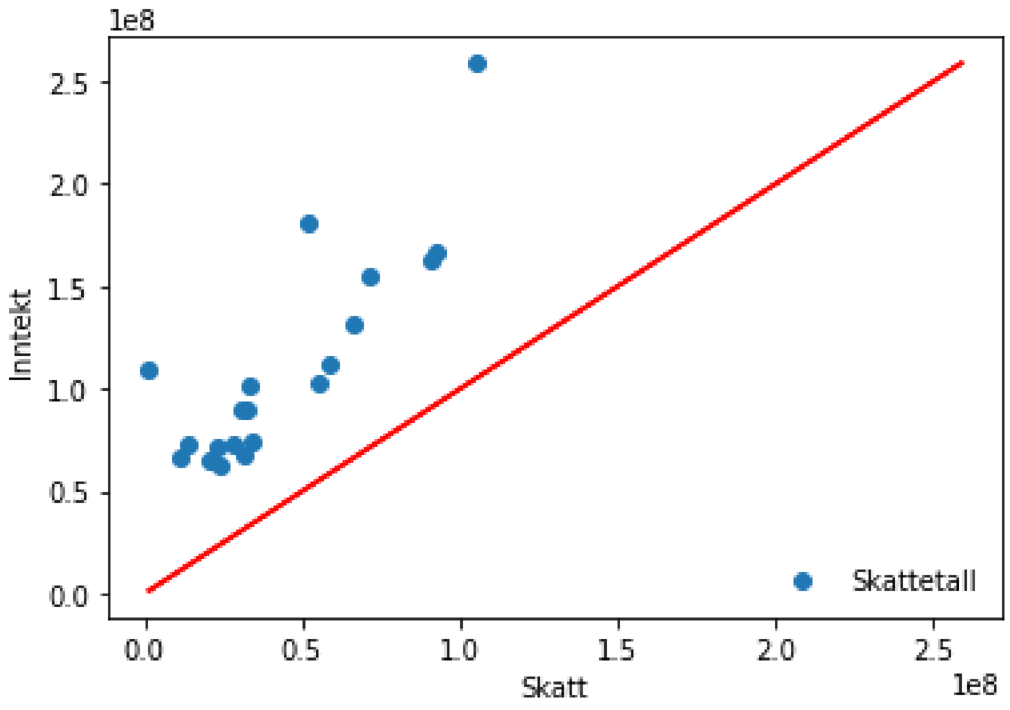
[12]:
```
res.params
```

[12]:
```
Skatt        1.000000e+00
Intercept    4.656613e-09
dtype: float64
```

[ ]:
```

```

[16]:
```
x=np.linspace(min((oslo['Skatt'])), max((oslo['Inntekt'])), 100)

regression_line=res.params['Intercept']+res.params['Skatt']*x

ax.plot(x, regression_line,color='red',label="Regression line")
fig
```

[16]:



[18]:
```
"Vi ser at regresjonslinjen ikke passer i plotten. Regresjon linja og antall␣
 ↪observasjoner som holder seg nærme linja vil si at det er sterk korrelasjon

"Sammarbeidet med Adrian Risberg, Andre Ydstebø, Mathias Hetland"
"Det meste av koden som er brukt/inspirert er fra Espen Sirnes"
"Litt usikker på hva som skjedde med regresjon linja mi"
```

[18]:
```
'Det meste av koden som er brukt er fra Espen Sirnes'
```