Assignment_1

September 14, 2022

1 Lab work 3

2 TASK 1

Calculating minimal SSR expression to fit a linear regression line

Instructions: Given a dataset of n values (x_i, y_i) for $1 \leq i \leq n$, describe the process to compute values α and β such that

$$\sum_{i=1}^{n}(y_i - \alpha x_i - \beta)^2$$

is minimal. This will yield a (linear) model $y = \alpha x + \beta$ adhering to the least-squares condition.

3 Process (in English):

1) Expansion of the brackets
2) Collect like terms
3) Obtain the 3 terms $x_i y_i$, x_i^2 and y_i^2
4) Having values for (x_i, y_i), these three terms can be plugged into the partial differentiation formulas for α and β to obtain their values for the minimal SSR
5) Plug values into $y = \alpha x + \beta$

or

1) calculate the sum of all xi, yi, xi^2 and $xiyi$
2) calculate α and β using their formulas
3) plug values into the linear regression equation $y = \alpha x + \beta$

4 TASK 2

5 Simple example

Given the dataset (1,3),(2,8),(3,6) and using the method described above, the calculation of the values of α and β for the minimal SSR gives:

$alpha$ = -182.3 $beta$ = 1.5

[ ]:
```
### TASK 3

## Implementing the algorithm in Python
```

```
# 1) calculate the sum of all xi, yi, xi^2 and xiyi, n = the total of␣
 ↪datapoints in the dataset

sum_xi = (xi1 + xi2 + ... xin)
sum_yi = (yi1 + yi2 + ... yin)
sum_xi^2 = (xi1^2 + xi2^2 + ... xin^2)
sum_xiyi = (xiyi1 + xiyi2 + ... xiyin)
n = (n)

# 2) calculate alpha and beta using their formulas and plugging the values for␣
 ↪xi, yi, yi^2 and xiyi

alpha = (n*sum_xiyi - sum_xi*sum_yi // n*sum_xi**2 - sum_xi**2)
beta = (sum_yi - alpha*sum_xi // n)

# 3) plug values into the linear regression equation y = alpha x + beta, the␣
 ↪formula can be drawn/completed using given datapoints

y = alpha*x + beta
```

[ ]:
```
## TASK 4

## Testing python implementation

# 1) Create example dataframe

# Import pandas library
import pandas as pd

# initialize list of lists
data = [[1,3], [2,8], [3,6]]

# Create the pandas DataFrame
df = pd.DataFrame(data, columns=[xi, yi])

# print dataframe.
df
```

[1]:
```
# 2) calculate \alpha and \beta using their formulas and plugging the values␣
 ↪for xi, yi, yi^2 and xiyi

sum_xi = (1 + 2 + 3)
sum_yi = (3 + 8 + 6)
sum_xi_2 = (1**2 + 2**2 + 3**2)
sum_xiyi = (1*3 + 2*8 + 3*6)
n = (3)
```

```
print(sum_xi)
print(sum_yi)
print(sum_xi_2)
print(sum_xiyi)
print(n)
```

```
6
17
14
37
3
```

[8]:
```
# 2) calculate alpha and beta using their formulas and plugging the values for␣
 ↪xi, yi, yi^2 and xiyi

alpha = (n*sum_xiyi - sum_xi*sum_yi // n*sum_xi**2 - sum_xi**2)
beta = (sum_yi - alpha*sum_xi // n)
print(alpha)
print(beta)

# The values for alpha and beta do not correspond with the ones previously␣
 ↪obtained and therefore might not be correct
```

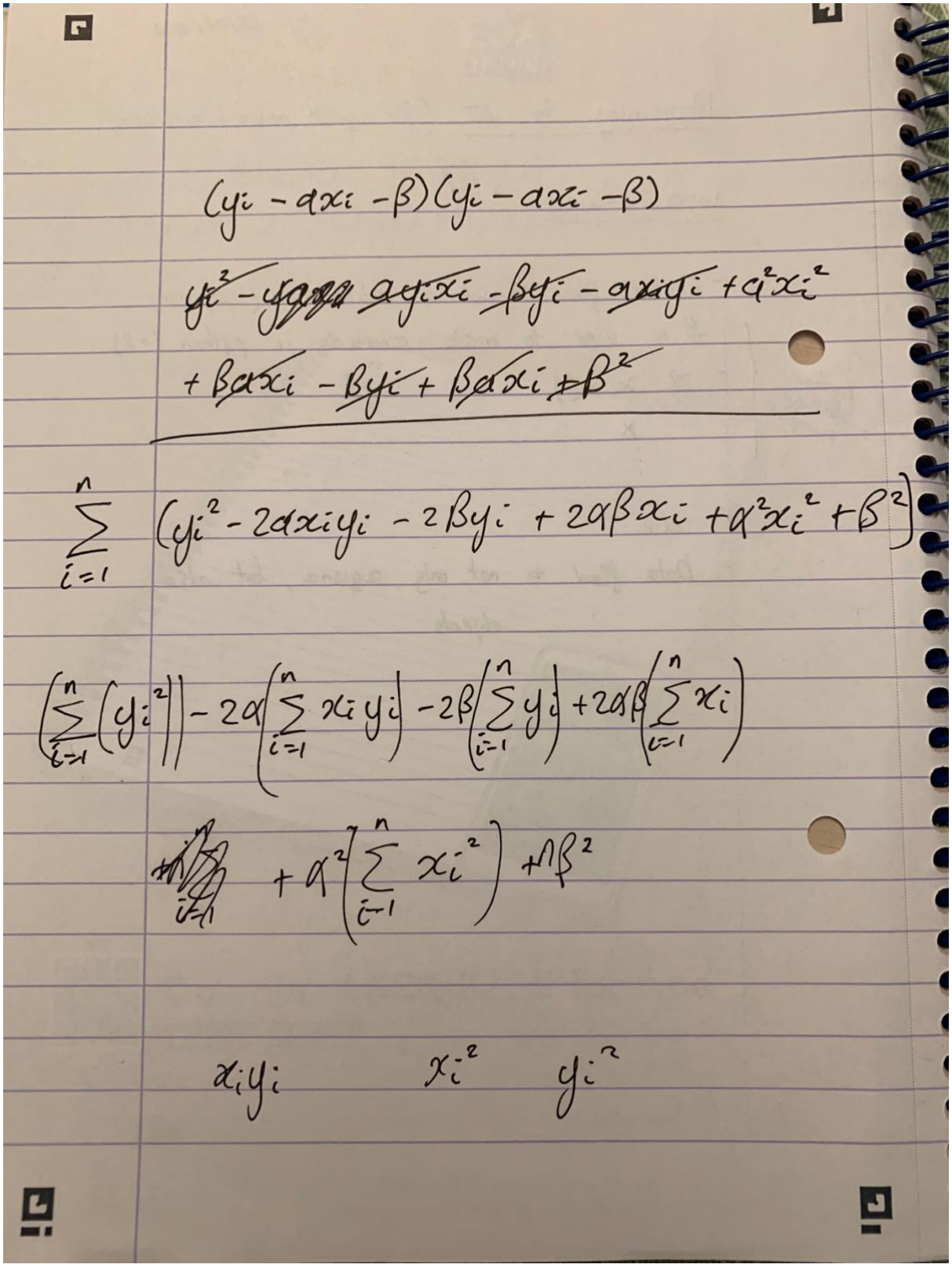
```
-1149
2315
```

[9]:
```
# 3) plug values into the linear regression equation y = alpha x + beta

# Example, finding the value of y for when x is 6 by using the obtained values␣
 ↪for alpha and beta

y = (125*6 - 233)
print(y)
```

```
517
```

[ ]:
```

```

CALCULATIONS



$$(y_i - \alpha x_i - \beta)(y_i - \alpha x_i - \beta)$$

$$y_i^2 - \alpha y_i x_i - \beta y_i - \alpha x_i y_i + \alpha^2 x_i^2 + \beta \alpha x_i - \beta y_i + \beta \alpha x_i + \beta^2$$

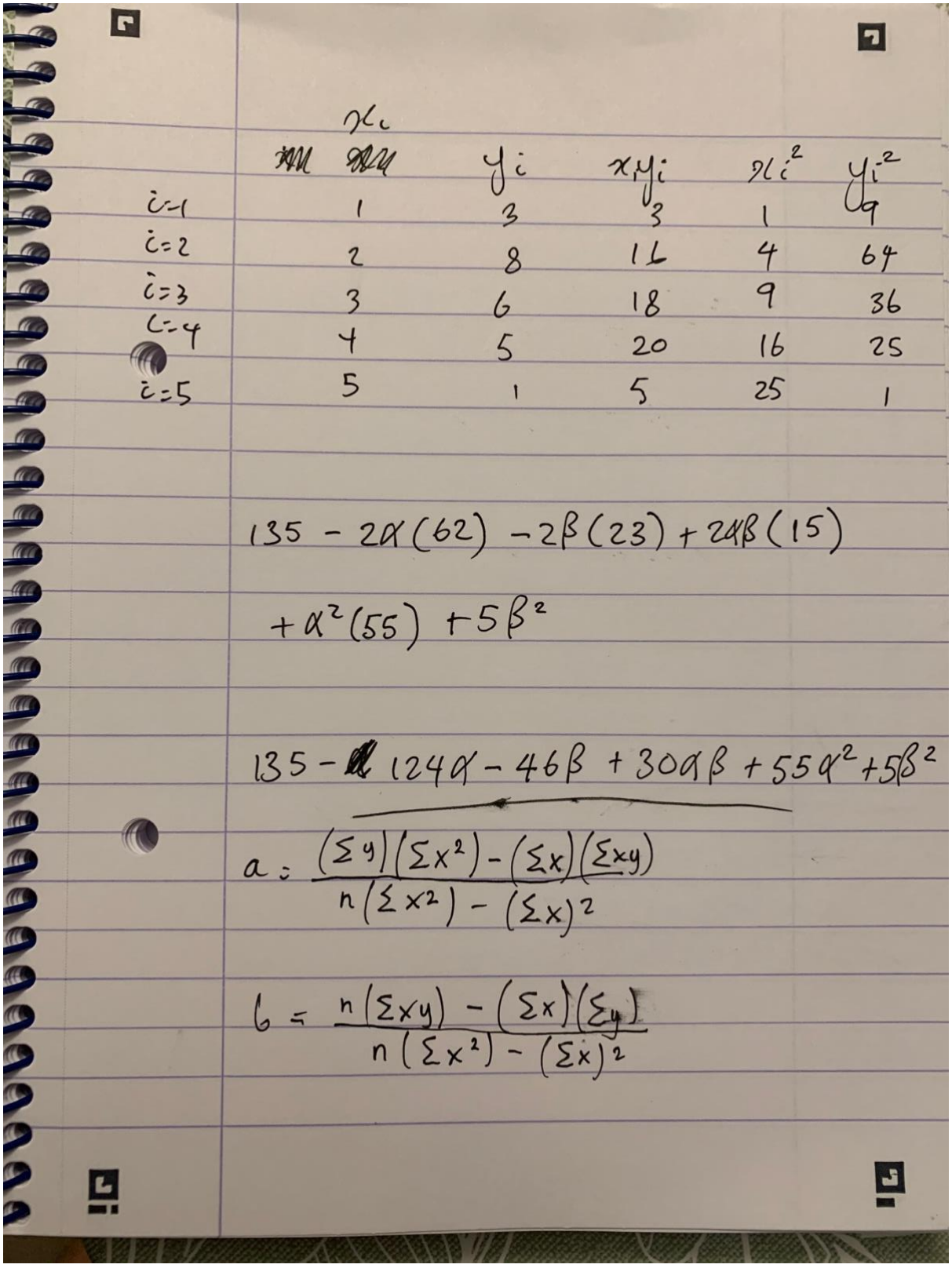
$$\sum_{i=1}^{n} \left(y_i^2 - 2\alpha x_i y_i - 2\beta y_i + 2\alpha\beta x_i + \alpha^2 x_i^2 + \beta^2\right)$$

$$\left(\sum_{i=1}^{n} y_i^2\right) - 2\alpha\left(\sum_{i=1}^{n} x_i y_i\right) - 2\beta\left(\sum_{i=1}^{n} y_i\right) + 2\alpha\beta\left(\sum_{i=1}^{n} x_i\right)$$

$$+ \alpha^2\left(\sum_{i=1}^{n} x_i^2\right) + n\beta^2$$

$x_i y_i$ $\quad x_i^2$ $\quad y_i^2$

CALCULATIONS



	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
$i=1$	1	3	3	1	9
$i=2$	2	8	16	4	64
$i=3$	3	6	18	9	36
$i=4$	4	5	20	16	25
$i=5$	5	1	5	25	1

$$135 - 2\alpha(62) - 2\beta(23) + 2\alpha\beta(15) + \alpha^2(55) + 5\beta^2$$

$$135 - 124\alpha - 46\beta + 30\alpha\beta + 55\alpha^2 + 5\beta^2$$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

CALCULATIONS

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
$i=1$	1	3	3	1	9
$i=2$	2	8	16	4	64
$i=3$	3	6	18	9	36
$\downarrow \Sigma$	6	17	37	14	109

$n = 3$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$\beta = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$a = \frac{17 \cdot 14 - 6 \cdot 37}{3 \cdot 14 - 6^2}$$

$$\frac{238 - 1332}{42 - 36} = \frac{-1094}{6} = -182.3$$

$$\beta = \frac{3 \cdot 37 - 6 \cdot 17}{3 \cdot 14 - 6^2} = \frac{111 - 102}{42 - 36} = \frac{9}{6}$$

$$= 1,5$$