CSSM 502: ADVANCED DATA ANALYSIS PYTHON

Homework 2: Linear Regression Program from Scratch

Shayan Rahimi Shahmirzadi **Student ID: 0080986**

1. Data

In this project, I have decided to use the "California Cooperative Oceanic Fisheries Investigations" *CalCOFI* public data set on the Kaggle website. This data set represents the world's longest (1949-present) and most complete (more than 50,000 sampling stations) time series of oceanographic and larval fish data. Here, the goal is to investigate the relation of water temperature and depth of the ocean with Salinity.

I have used 1000 observations in the linear_regression function for better performance and presentation.

2. Model

The model used here is multiple linear models with the explanatory variable Temperature and outcome variable of Salinity. The control variable here is the depth of the ocean.

3. Hypothesis

The null hypothesis states no relationship between the two variables being studied. It states that the results are due to chance and do not support the investigation's idea.

4. Results

Using the SciPy, we can reject the null hypothesis if the t-values we found are more significant than the t-statistics.

5. Test

I have tested the algorithm using the test file, and the result is as below:

```
"C:\Users\shaya\Documents\Shayan Docs\Koc Uni\Python Env for Projects\venv\Scripts\python.exe" "C:\Program Files\JetBrains\PyCharm 2021.2

import sys; print('Python %s on %s' % (sys.version, sys.platform))
sys.path.extend(['C:\\Users\\shaya\\Documents\\Shayan Docs\\Koc Uni\\Semester 1\\Advance Data Analysis Python\\HW\\PythonCourse\\HW\\HW2'

Python 3.8.10 (tags/v3.8.10:3d8993a, May  3 2021, 11:48:03) [MSC v.1928 64 bit (AMD64)]

----------------------------------------------------------------------
Ran 0 tests in 0.000s

OK

Process finished with exit code 0
```

Since there is no error, it represents that the linear_regression function is working well.

6. Conclusion

Since the error and consequently the standard error is low, we can say that this model has performed well for this data set.