Tahsin Alam

CS 21700

Project 2

Introduction:

From this report we expected to find the positive correlation. We are trying to see whether our reports value does predict and match the values and our all the findings that we have in our original findings form our statistical part and from here we should state that all our findings should correlate to each other the practical results we are expected to see here. The practical reason behind this we should be able to figure out the distribution and the cortrelation between their efficiency.

Population:

Population size: First stage we are defining our population sample 100 to get corresponding predictors and responses.

Sample size: We are picking every 10^{th} sample from our population size, defining our sample size to be 10.

Variables:

Hash String: Tahsin= `56A78E6D6F05A68686459468895C8F06`
Four parts:

```
56A78E6D
6F05A686
86459468
895C8F06
```

intercept β_0 = `0.6769884143383188`
slope β_1 = `0.8673599394351988`
Standard Deviation: `1.073137167409592`
Predictors:

```
xs=Predictors: [ 0.75935688  0.97164309  0.60174187  0.1628279   0.4338501
  0.81683472  0.52290488  0.66458728  0.78740283  0.68739841  0.54361862
  0.76734154  0.96235792  0.45685664  0.40571089  0.54112995  0.01174162
  0.46868921  0.79117279  0.62558308  0.04021772  0.21693604  0.84560746
  0.76939935  0.15797588  0.07150331  0.34856132  0.71756574  0.93960449
  0.18565904  0.89088097  0.72221526  0.01223057  0.99097945  0.67485857
  0.22088049  0.34685693  0.17292248  0.0364026   0.52266691  0.97199479
  0.31873483  0.48406105  0.59747196  0.07811448  0.15622616  0.96906549
  0.17265809  0.0184472   0.42111586  0.12098735  0.35437953  0.78651734
  0.72944772  0.00598579  0.27213558  0.05143405  0.30285406  0.55467052
  0.14180774  0.89852584  0.17595     0.37775609  0.42288997  0.8596258
  0.65514684  0.86712788  0.75518729  0.67500925  0.37407115  0.87393565
  0.16670983  0.35913438  0.48904314  0.09069532  0.61117955  0.35040709
```

```
  0.78961583  0.26612543  0.46553045  0.09451631  0.75608351  0.85058119
  0.48812344  0.0951247   0.10733605  0.35859052  0.30261445  0.12269738
  0.94079119  0.99474843  0.6278918   0.95656005  0.49215686  0.45176754
  0.85369707  0.36960166  0.94776158  0.98955026]
```

Responses:

Ys=

```
Responses: [ 2.07094931  1.61922873  0.86987869  0.32117058  1.17874938 -
0.02217831
  1.05671391  1.9755661   2.35522792  0.97797911  1.95344458  0.19873833
  1.08381693  1.14815062  3.61579618 -0.70652771 -0.49027333 -0.03103092
  2.17745506  1.81663628  1.69580879  0.51706064 -1.07879235  2.22038337
 -0.33573748  2.46587546 -0.70616864  0.50062038  2.39433455  1.15236439
  1.19631347 -0.17733284  0.99339252  0.11925699  1.64937987  0.95873171
 -0.02396834 -0.69968876  0.46605829  0.29553659  2.59189879  0.72608866
  2.06598864  1.92486304  1.34650551 -0.76846542  0.96671692  0.82316727
 -0.70424326  0.37734207  2.16280926 -1.07746155  0.1153737  -1.71657019
  1.19959151  0.56219703  0.13334967  0.93947586  3.13843819  0.40933563
  1.0050806   1.96382097  2.93747544  2.88167674  1.69435937  1.37912194
  0.65934293  2.27417184 -0.03179414  0.30437726  3.68707078  2.54391203
  2.46599842  0.14245843  0.5925456   0.22717716  1.69950686  1.61308617
  1.97338654  0.46577359  1.07740173  0.46478385  0.58726341  1.44339971
  0.73023105  1.5286837   0.88242948  0.7661545   1.38712257  0.10897747
  1.80431789  1.30129639 -1.18710465  0.8527679   0.37639552 -0.04301314
  2.79068322  0.55452118  0.67606681  1.62913826]
```

X= Picking every 10th value from xs

```
0.78740283  0.79117279  0.93960449  0.0364026   0.0184472   0.55467052
  0.67500925  0.26612543  0.12269738  0.98955026
```

Y= picking every 10th value from ys

```
0.60481918  0.23033616  1.11636605  0.66399585  0.58358083  1.11207168
  0.52614468  0.70000798  2.59548246  0.49788883
```

At this point we will define variables we used for out Anova table, Anova F test, CI for the slope and CI for the mean of the responses

Variables for Anova F table:

1. Mean of x: xm= 0.518108274886
2. Mean of y: ym= 0.863069367875
3. Regression Sxx: 1.27450178913

$$S_{xx} = \sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

4. Regression Syy: 3.98566913649

$$S_{yy} = \Sigma(y - \bar{y})^2$$

5. Regression Sxy: -0.759861868273

$$S_{xy} = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

6. Total sum of squares: SS_TOT= 3.98566913649

$$SS_{TOT} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)s_y^2$$

7. Regression sum of squares: SS_REG= 0.958824601636

$$SS_{REG} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

8. Error sum of squares: SS_ERR= 3.02684453485

$$SS_{ERR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

9. df_TOT = n-1=9
10. df_REG=1
11. df_ERR=df_TOT-df_REG=8
12. Mean square: MSR=0.958824601636
    MSR=SS_REG/df_REG
13. Mean square error: MSE= 0.378355566856
    MSE=SS_ERR/df_ERR
14. F= 2.53418922735

$$F = \frac{MS_{\text{REG}}}{MS_{\text{ERR}}}$$

Variables used for Anova F test:

15. k=1
16. R square= 0.240568037336

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}}$$

Anova F $= (R^2/k)/(1 - R\text{^}2)/(n - k - 1)$

Variables used for Confidence Interval:

17. α = 0.05 # confidence level
18. ta= 2.306 # Critical t score from the t distribution table.
19. Regression variance: s2= MSE
20. σ_b1= 0.544853618377
21. CI for the slope: CI_b1 = (b1-ta*σ_b1, b1+ta*σ_b1)
    (-0.3890725045412774, 2.1237923834116752)

Variables used for CI for the mean of the responses:

22. xe=2, ye= b0+b1*xe= 2.4117082932087164
23. σ_μe= 0.830513717316
24. ME_ye = ta*σ_μe = 1.91516463213
25. CI_ye= (0.49654366107756776, 4.326872925339865)

Study Design:

Starting from the first we generate the MD5 hash with the designated string and the 32-bit number was divided in four parts to perform different operation on each of them to determine slope, intercept, standard deviation and the random generator seed. We choose our population size to be 100 and from there we figure out the corresponding 100 responses and predictors by using the random seed value and with python programming. We construct a linear regression model with scatter plot and the fit linear regression line. We use xs and ys as our sample variables for

population from where we construct the first linear regression with scatter plot. The function G(xs) was used to construct the plot.
In the second part we got the value x and y from taking every 10^{th} sample from xs and ys. The second graph was constructed by taking the x and y as our sample. The value G(x) was used which was declared first. The sample linear regression model was constructed using x and y as our experimental variable and G(x) .

***Results descriptive analysis*:**

sample size	Mean(x)	Mean (y)	Std Deviation
10	0.518108275	0.860693679	
100	0.49737793	1.001993868	1.07313716

Fig 1: Sample mean, Standard deviation, and mean.

Mean (x)	5 Number Summary
25 percentile	0.863069368
50 percentile	0.614839886
75 percentile	0.790230297
max	0.989550262
min	0.184471925

Fig 2: 5 number summaries for Mean of x.

Mean (y)	5 Number Summary
25 percentile	0.158554395
50 percentile	0.614839886
75 percentile	0.790230297
max	0.863069368
min	0.018447195

Fig 3: 5 number Summary for mean of y.

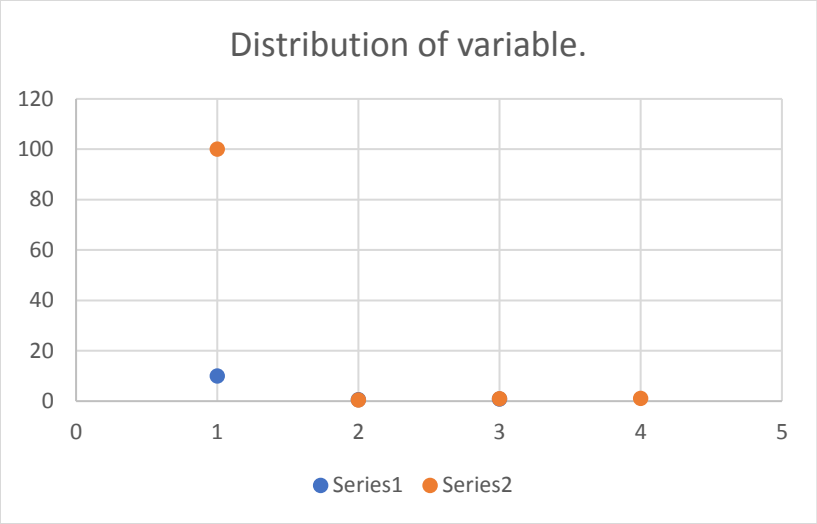

Fig 4: Boxplot distribution.

All the tables and charts show the mean of x and y and the standard deviation and a chart for 5 number summaries. Distribution of the variable shows how each of variables represent in the boxplot and their position.

Results statistical analysis:

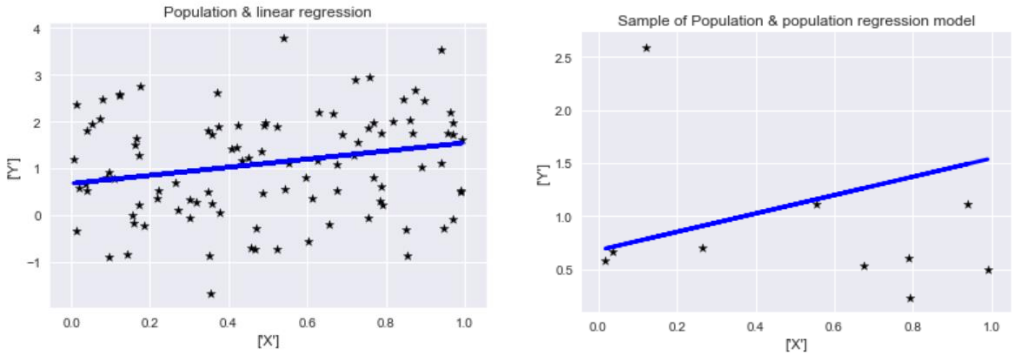

Fig 5: Population and sample linear regression model.

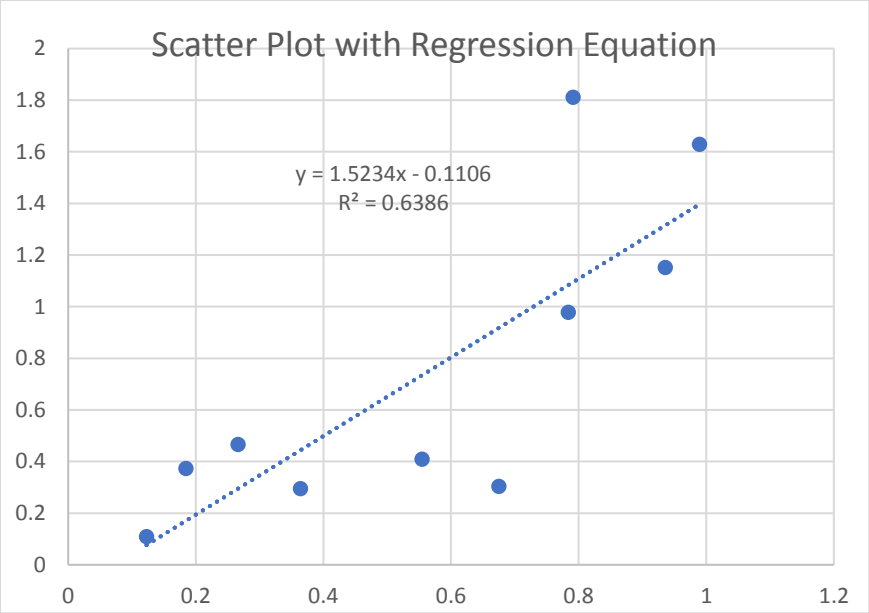

Fig 6: Scatter plot with x and y value to show equation and r square value.

Source	SS	DF	MS	F-score
Model	3.985566	9	0.95882462.5341489	
Error	0.95888824	1	0.37835557	-
Total	3.026844534	8	-	-

Fig 7: Anova Table.

Confidence Intervals: (-0.389072504412774, 2.12379234116572) for the slope.

Findings:

After findings all the statistical analysis and the results of all the corresponding equation we can see that context support our original research questions and it also support our original findings from the programming part that we found and the analysis part here that we have discussed match our analysis.

Discussion:

If we can draw any conclusions from there we can say that all our findings from the original part and the analysis part do support our results of the study and the results that we expected.
If there were any factors that might affect are the experiments are the mean of x and y and the Variables we used to find them and the standard deviation. If we learn anything from the project the project that should be working with the sample and population size and finding the corresponding results and the chart that shows the distribution of each variable. The research variable also plays a major part to conclude our findings and to show the distribution.
The statistical analysis we conducted also support our findings and justify all the equations and values that we predicted before which gave us a better understanding.