DS 203 E2

Anupam Vinay Singh

24B2120

Note:- I wasn’t able to do good in this assignment. I know I have done lots of error and mistakes but please do not consider it as fraudulent submission because I am submitting whatever I can make so that I can get the report of my work. I haven’t copied and tried on my own.

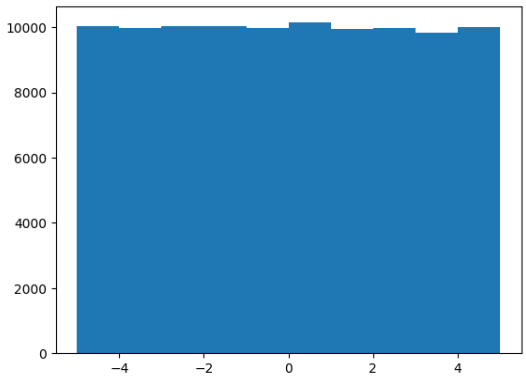
Part 1

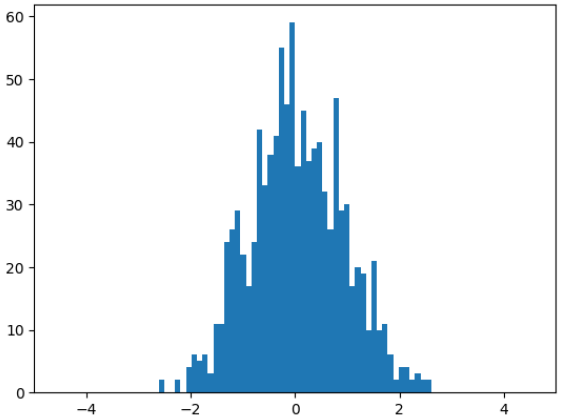
(a) standard deviation of the sampling distribution = population standard deviation/root(size of population).

(b) When we use different distribution models, the shape of population changes. The sampling distribution of the mean follows central limit theorem. Following are changes for all the models:-

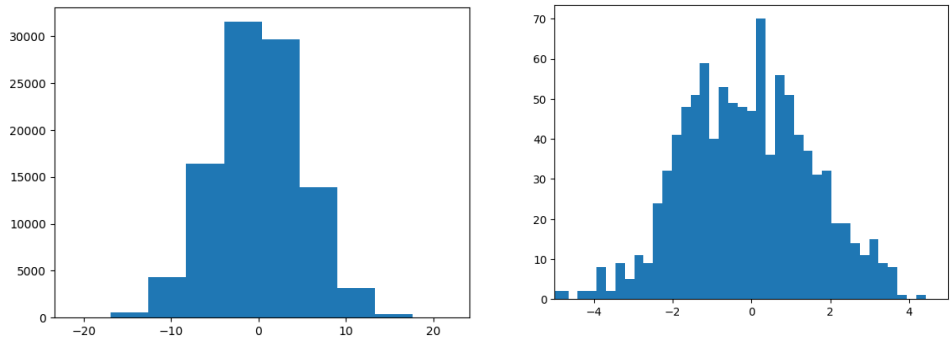
```
pop = np.random.uniform(-5,5,pop_size):- population shape is flat and regular, sampling distribution of the ,mean is bel shaped.
pop = np.random.normal(0, 5, pop_size):- both are normal, smooth bell shaped.
pop = np.random.poisson(5,pop_size):- population shape is right sckewed, sampling distribution of mean is skewed for small n
pop = np.random.binomial(7,0.7,pop_size):- discrete and right skewed, smoother with larger n.
pop = np.random.triangular(1,3,5,pop_size):- bell shaped and right skewed, smooth bell shaped even at moderate n.
```

(i) `pop = np.random.uniform(-5,5,pop_size):-`

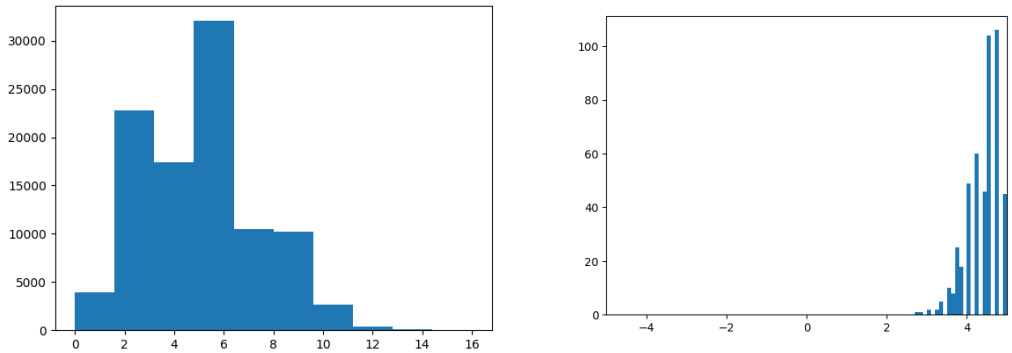




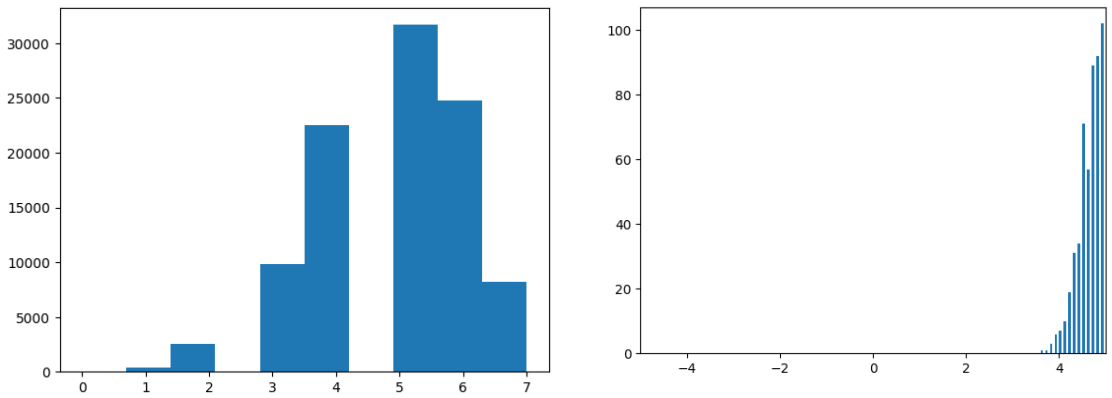
(ii) `pop = np.random.normal(0, 5, pop_size):-`



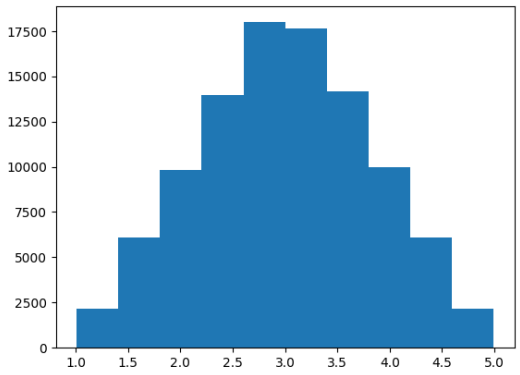
(iii) `pop = np.random.poisson(5,pop_size)`

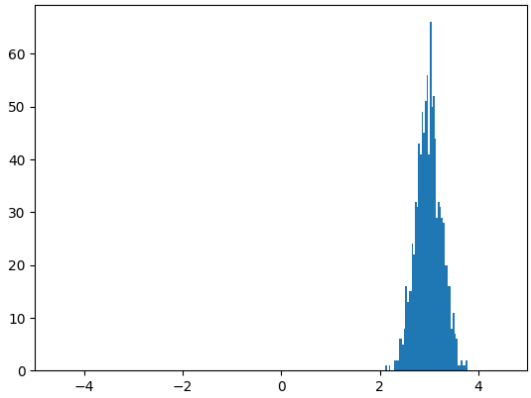


(iv) `pop = np.random.binomial(7,0.7,pop_size)`



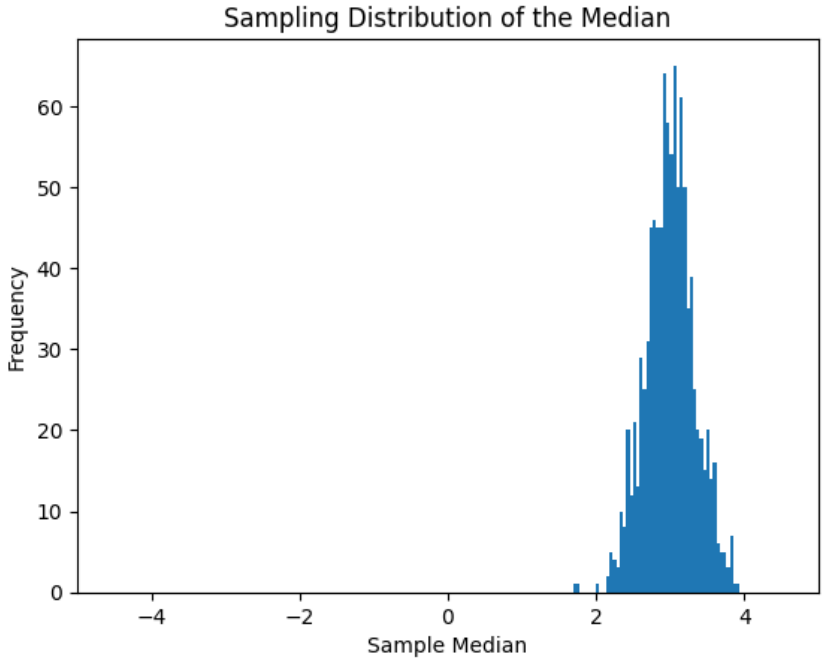
(v) `pop = np.random.triangular(1,3,5,pop_size)`





(c)When we increase the population size, the mean come more closer to 0. The sampling distribution becomes more narrow and concentrated near the true population mean. It becomes more normal.

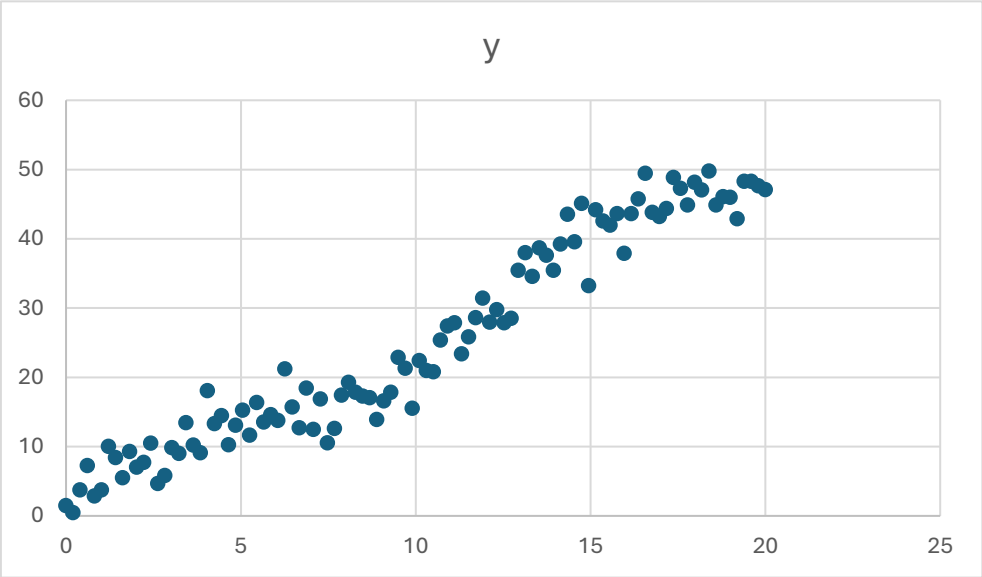
(d) Added the sampling distribution of median. The median is more wider than the mean and less normal. We can conclude that mean is more efficient.



This is the code after addition :-
https://colab.research.google.com/drive/1DwQ7pueH5yRx4xkpJXgk_lPbXHvcAtbJ?usp=drive_link

Part 2

(a)The data shows an approximate linear trend.



Conclusion:- We can take it in the linear regression model.

(b) Pearson's constant = r= Cov(X,Y)/ σXσY

Where Cov(X,Y) = covariance between X and Y and

σX , σY = standard deviations of X and Y

But here we didn't calculated it directly y thus formula.

Rather we took mean x, mean y, calculated deviation in x and y, computed product of deviations, computed squares of deviations, and then applied Pearson's formula:-

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}$$

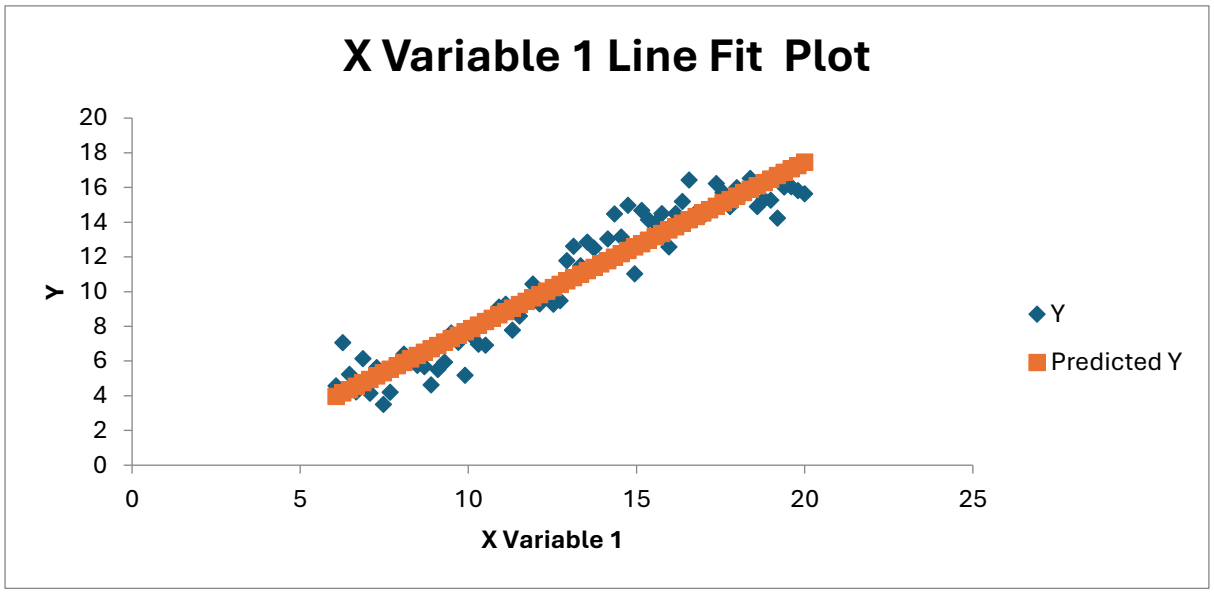
(c)

Column1	ysd5	ysd10	y	ysd20	ysd25
R^2	-1.55701	0.708939	-1.55701	-0.49418	-0.19969
p-val(intrcpt	0.000251	0.000251	0.000251	0.018569	0.952714
p_val(x)	6.43E-38	6.43E-38	6.43E-38	1.9E-11	0.000291
F-statistic	720.9735	720.9735	720.9735	64.53847	14.59715
RMSE	2.365097	1.595896	1.595896	18.53208	17.72289

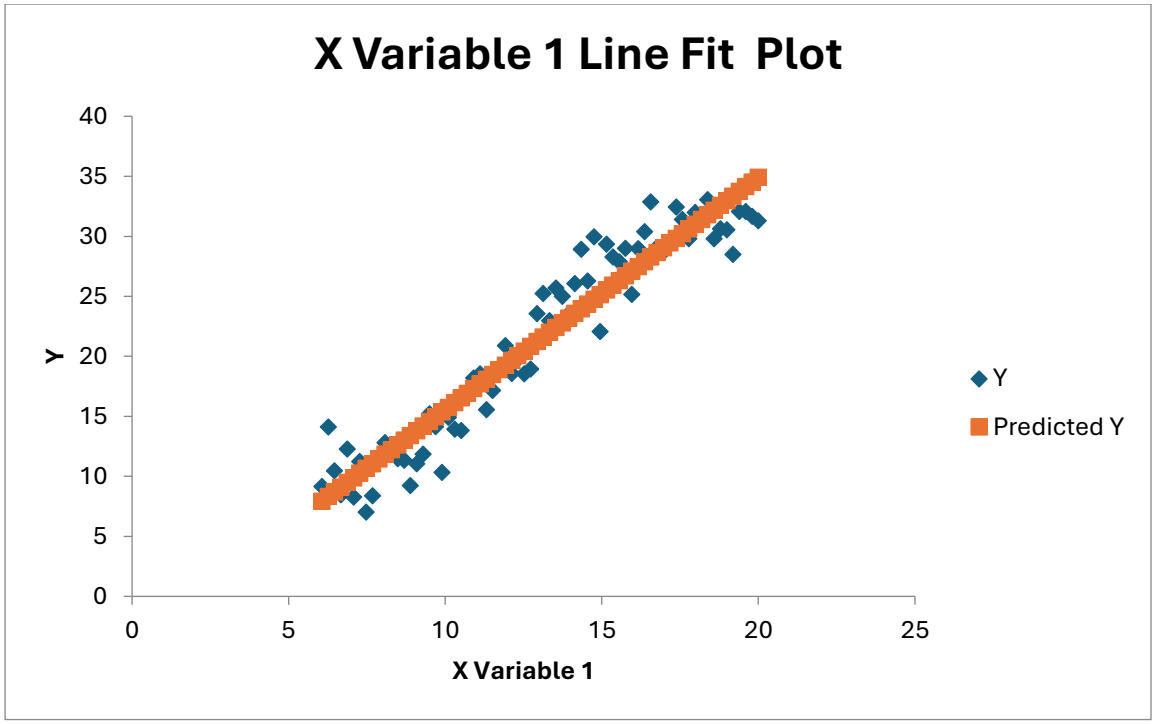
(Note:- R^2 can not be negative. I took the first 30 values as test and the rest of 70 values as train. This error probably came because the initial values were outliers.)

The plots:-

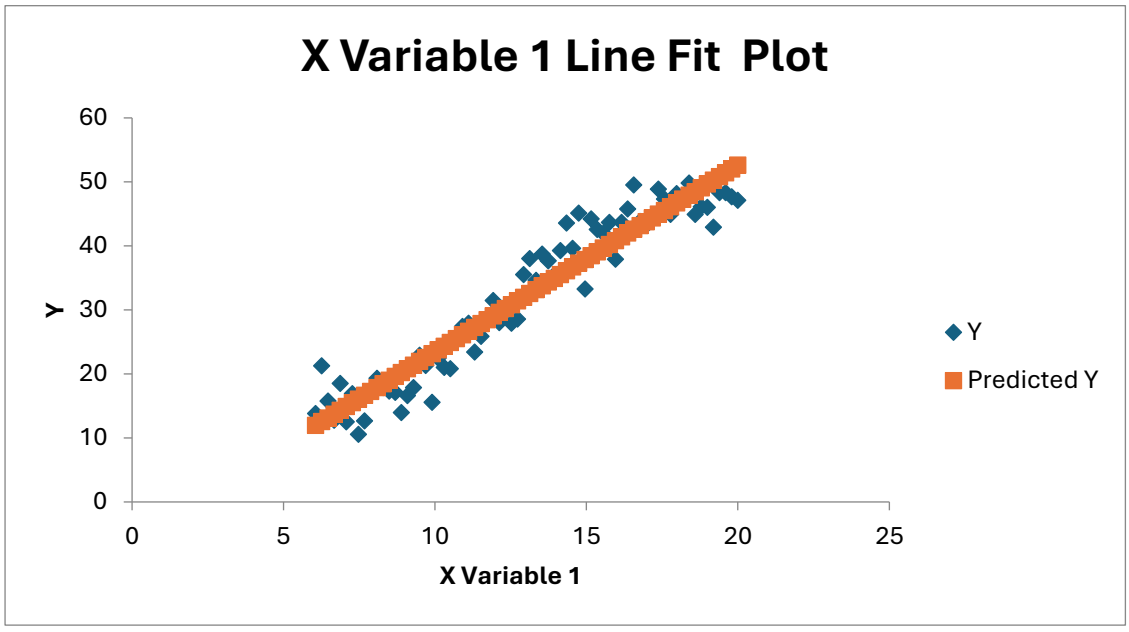
Ysd5:-



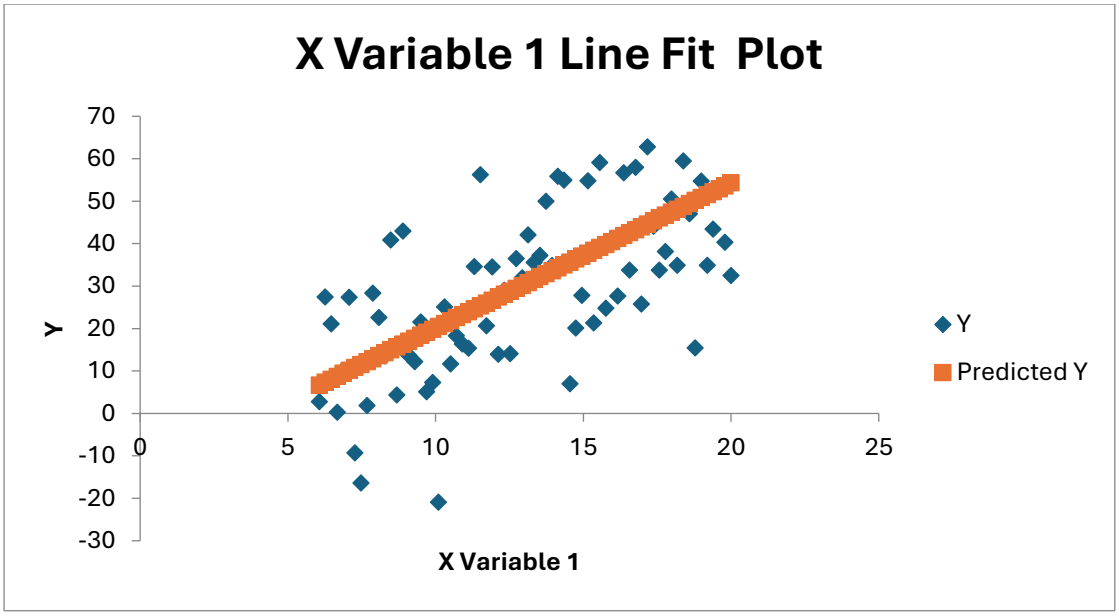
Ysd10:-


X Variable 1 Line Fit Plot
40
35
30
25
20
15
10
5
0
Y
0
5
10
15
20
25
X Variable 1
Y
Predicted Y

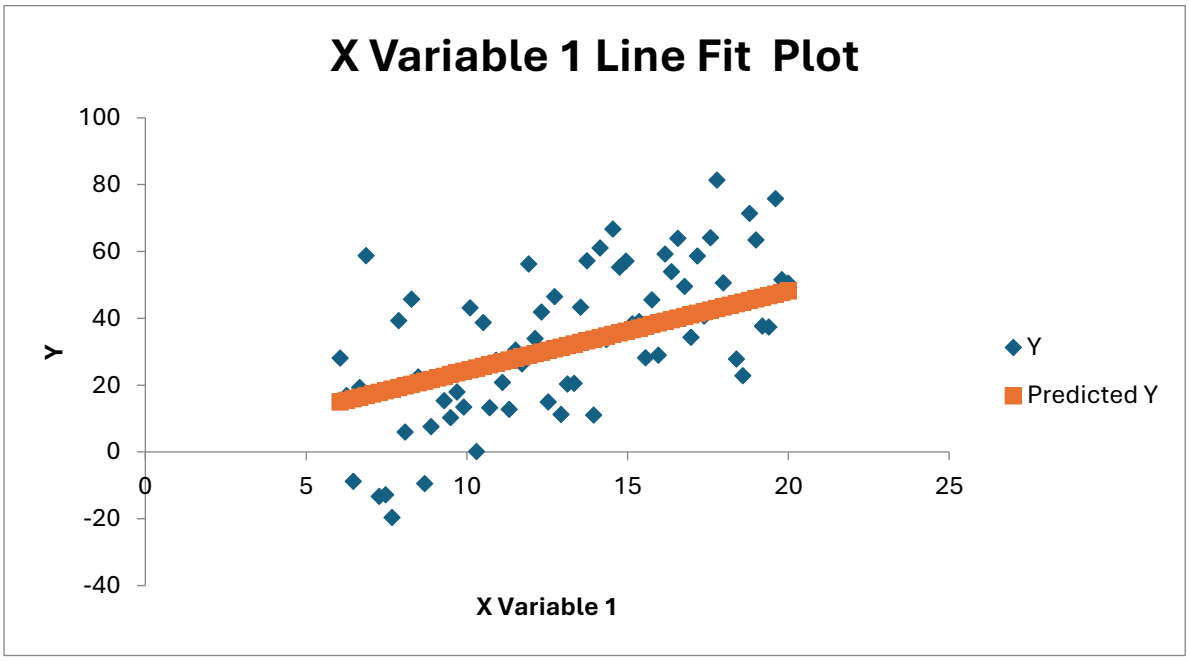
Y:-


X Variable 1 Line Fit Plot
60
50
40
30
20
10
0
Y
0
5
10
15
20
25
X Variable 1
Y
Predicted Y

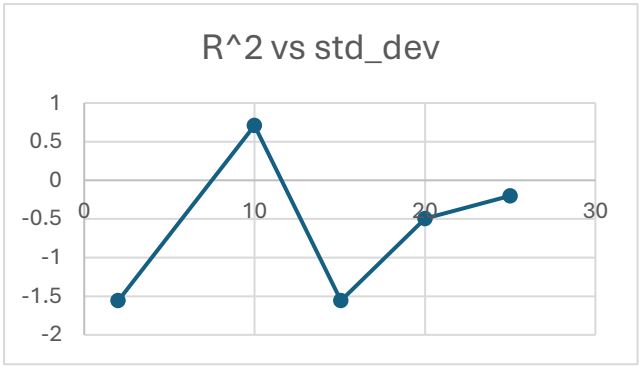
Ysd20:-

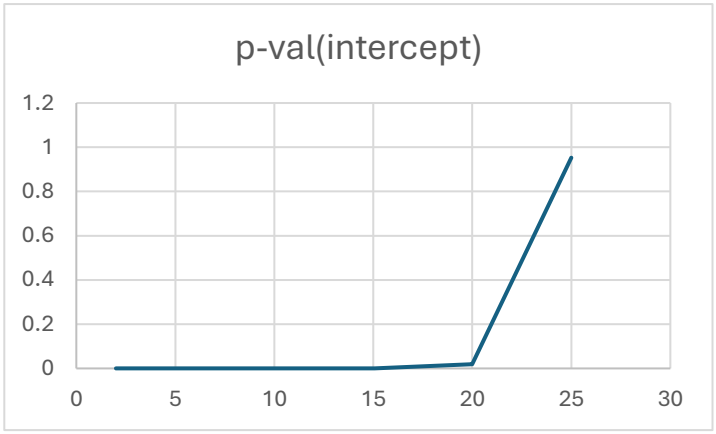


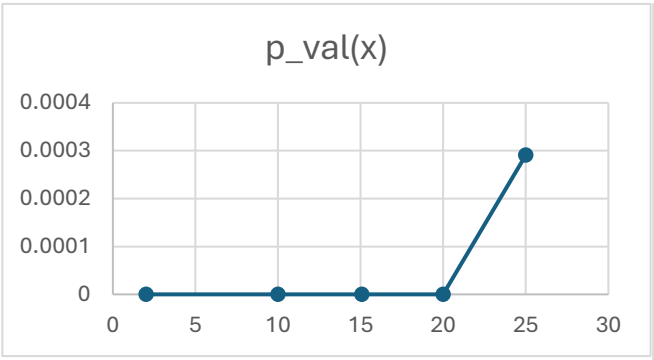
Ysd25:-

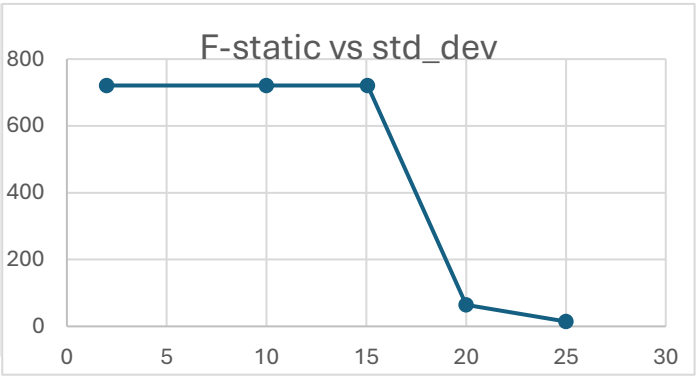


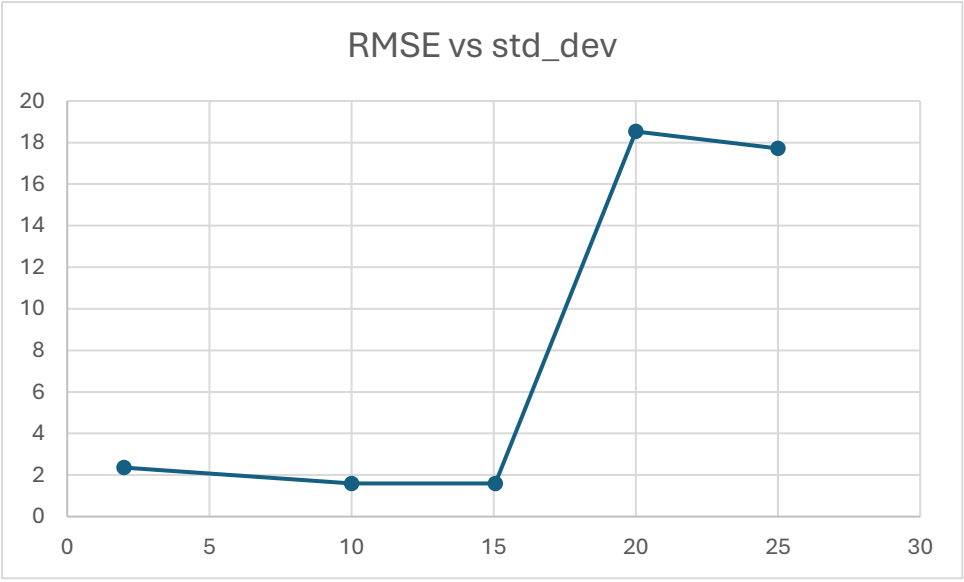
(d) All the graphs of metrics vs standard deviations:-











The p- values and the RMSE increases as we increase the standard deviation. P- value (intercept) increases because it is harder to estimate intercept. P- value (x) increases because of more noise. RMSE increases because of larger residuals due to more noise. F static decreases as we increase the standard deviation because signal to noise ratio decreases. Ideally R^2

should decrease as the model explains less variance but it is not visible on my plot because of some inaccuracies.

(e)Std_dev 5:- CI: (-2.88, -0.92) and (0.89, 1.04)
The intervals are very tight. This means our model is precise.
Std_dev 10:- CI: (-5.77, -1.84) and (1.79, 2.08)
Actual:- CI: (-8.69, -2.77) and (2.69, 3.13)
Std_dev 20:- CI: (-16.51, 17.53) and (1.14, 3.63)
Interpretation:- At low std, CI are narrow that means high precision. At high std, CI widens decreasing the precision and increasing uncertainty.