DS 203: Programming for Data Science - Exercise 2

Swayam Saroj Patel (22B1816)

February 6, 2025

Question 1:

Part a:

Scatter plot:

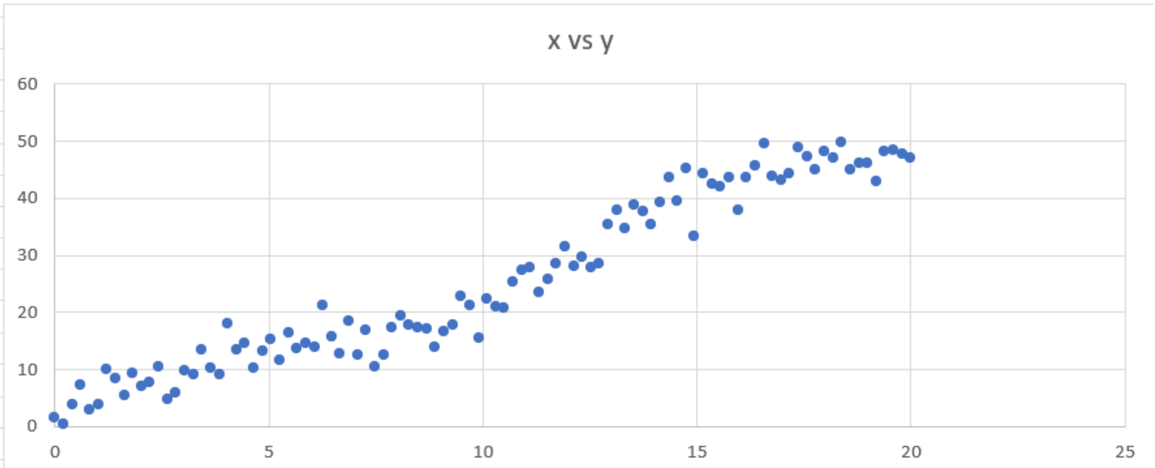


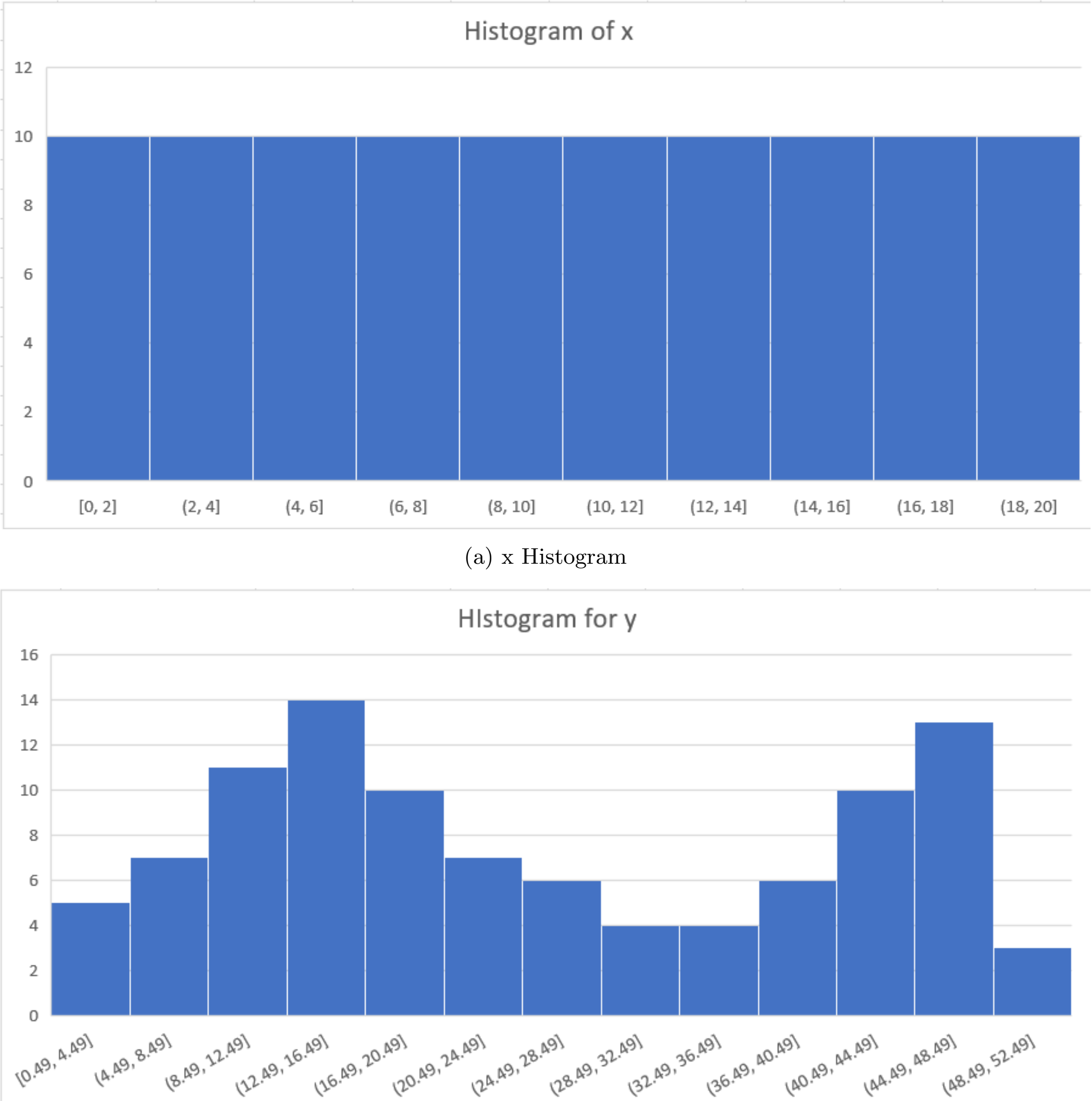
Figure 1: Scatter plot

The scatter plot reveals a positive slope linear like behavior of the data. It seems that our normal simple linear regression can predict well here.

There seems to be some deviation in the middle and at the end it again curves up but I think it won't cause too much of an error.

Histograms:

The histogram of x follows a uniform distribution, meaning values of x are evenly distributed across their range.
The histogram of y shows an irregular pattern with peaks at certain intervals, suggesting that y is not normally distributed but still roughly symmetric.



(a) x Histogram

(b) y Histogram

Figure 2

Part b:

The **SLR** which ha the model function as $\mathbf{y = mx + c}$ seems to be a good choice here. The reason is because the scatter plot shows a linear behavior. So it is highly feasible to use SLR.

The only hiccup one may say that could cause some usual finding s is that even if x is uniformly distributed y is not. That is is the data follows a linear model perfectly then the values in y too should have a roughly uniform distribution, that is the distribution of y and x should be similar.

Part c:

Metric	**Value**
Mean of x ($\bar{x}$)	10
Mean of y ($\bar{y}$)	25.4052
Median of y	21.2824
Variance of y	224.5781
Standard Deviation of y	14.9859
Range of y	49.2949
Skewness of y	0.2174
Kurtosis of y	-1.3960

Table 1: Descriptive Statistics for x and y

- The lower median of y than its mean indicates some right skewness.
- The variance which is not that less than the mean also indicates a wide spread in data which is verified by the almost double spread (maxy-miny) of the data
- A small positive skew implies y has a slight right inclination, but it's close to symmetric.
- A negative kurtosis suggests a platykurtic (flat) distribution, meaning fewer extreme values compared to a normal distribution

Expected Quality of SLR:

- The scatter plot indicates a **positive linear trend**, meaning y increases as x increases. Since no significant curvature is observed, a **linear model is appropriate** for capturing the relationship.
- The **high variance (224.5781) and standard deviation (14.9859)** indicate a **large spread** in y. This suggests that individual predictions may have some variability.
- The near zero skewness and negative kurtosis indicate that y follows a well-behaved distribution, supporting a **stable regression model**.

Given the **linear trend** and well-behaved distribution of y, an **SLR model should perform well** in predicting y

Question 2:

Part a:

The Pearson Correlation Coefficient between x and y is **0.9674**. This value is very close to 1, indicating a strong positive linear relationship between x and y. This suggests that as x increases, y also increases nearly linearly.

Part b:

After applying the logarithmic transformation to y (z = log y) and square root to x (w = sqrt(x)), the Pearson Correlation Coefficient between w and z is **0.9390**
This value is still high but slightly lower than that between x and y, meaning that the linear relationship has slightly weakened after transformation.

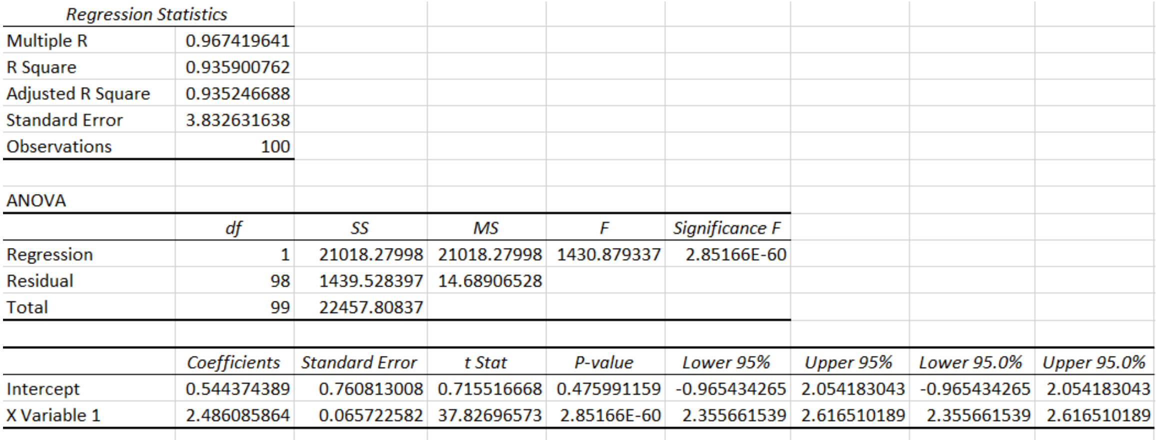
Correlation	**Value**
$r(x, y)$	0.9674
$r(w, z)$	0.9390

Table 2: Comparison of Pearson Correlation Coefficients

- The transformation has **weakened** the relation between the data.
- The operation log compresses higher values and so does square root but here both the compression aren't compressing the data equally, so the already highly linear pattern is deviated and the correlation decreases.
- A log transformation is beneficial if the relation between x and y is **exponential** that is there is some power relation between the two.

Question 3:

Part a:



Regression Statistics	
Multiple R	0.967419641
R Square	0.935900762
Adjusted R Square	0.935246688
Standard Error	3.832631638
Observations	100

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	21018.27998	21018.27998	1430.879337	2.85166E-60
Residual	98	1439.528397	14.68906528		
Total	99	22457.80837			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.544374389	0.760813008	0.715516668	0.475991159	-0.965434265	2.054183043	-0.965434265	2.054183043
X Variable 1	2.486085864	0.065722582	37.82696573	2.85166E-60	2.355661539	2.616510189	2.355661539	2.616510189

Figure 3: Regression results

We get the above result by using the formula $\mathbf{y = a + bx}$

Part b:

The model shows a high goodness-of-fit with an R^2 value of 0.9359. This tells us the model predicts the y value very well. The adjusted R^2 is 0.9352, confirming the strong explanatory power of the model.
The F-statistic of 1430.88 with a p-value of 0.47 and 2.85 x 10^{-60} strongly indicates that the model is statistically significant.

Part c:

Set	**Mean Value**
Predicted y	25.40523303
Actual y	25.40523303

Table 3: Comparison Mean Values

Since the predicted mean and the actual mean of y are the same, it indicates that the linear model has a strong predictive accuracy for the overall dataset.

Part d:

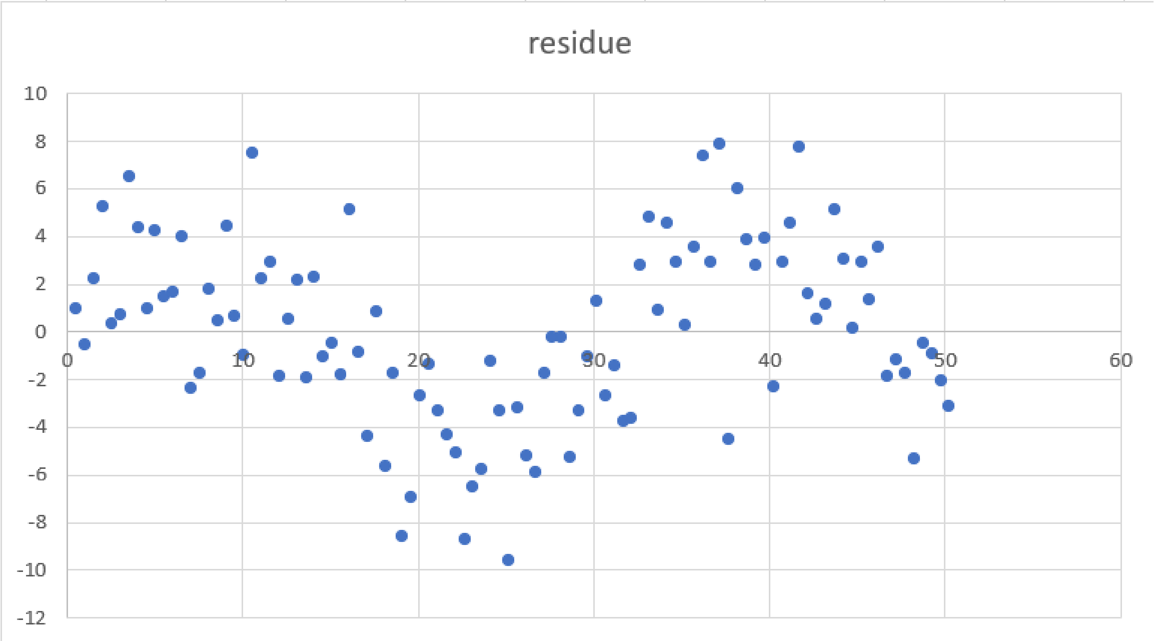


Figure 4: Residue vs Fitted values plot

The residue plot looks a bit like a sinusoidal wave. This suggests that there may be some small non-linearity in the relationship between the predictor and response variable. This indicates that a simple linear model might not fully capture the underlying trend in the data

Question 4:

Part a:

For standard deviation = 5 :

Table 4

Metric	Value
R^2	0.9359007
p-value (Intercept)	0.4759911
p-value (X Variable)	2.85165 x 10^{-60}
F-statistic	1430.8793
RMSE	19.6546549

For standard deviation = 10 :

Table 5

Metric	Value
R^2	0.9359007
p-value (Intercept)	0.4759911
p-value (X Variable)	2.851659 x 10^{-60}
F-statistic	1430.879337
RMSE	9.81348045

For standard deviation = 20 :

Table 6

Metric	Value
R^2	0.3359183
p-value (Intercept)	0.00136084
p-value (X Variable)	2.646963 x 10^{-10}
F-statistic	49.57222
RMSE	12.83102549

For standard deviation = 25 :

Table 7

Metric	Value
R^2	0.237988
p-value (Intercept)	0.0629233
p-value (X Variable)	2.63089 x 10^{-7}
F-statistic	30.6069
RMSE	20.70516495

Part b:

As the standard deviation increases, we observe the following trends:

- R^2 decreases, indicating a weaker model fit as spread or noise increases.
- The F-statistic decreases, reflecting a reduction in the overall explanatory power of the model.
- RMSE fluctuates but tends to increase, signifying greater prediction errors.

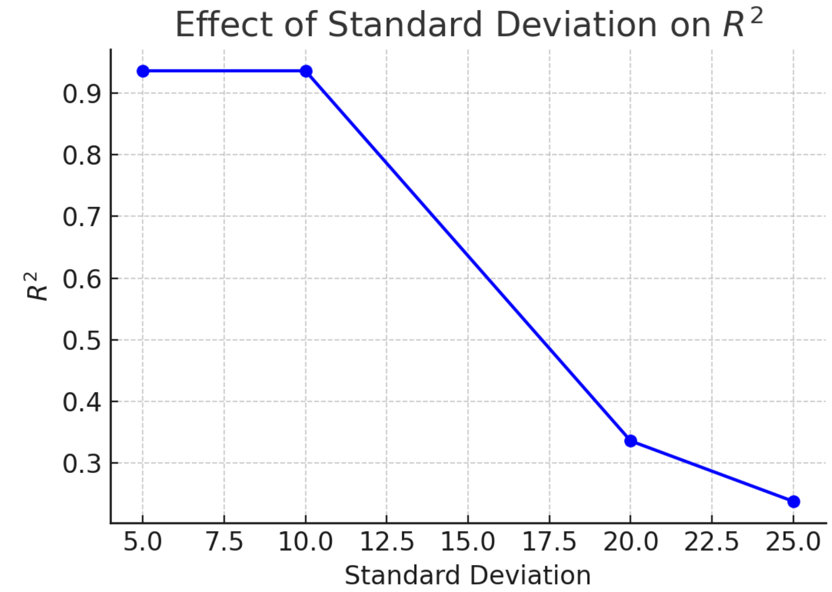


Figure 5: Effect of Standard Deviation on R^2

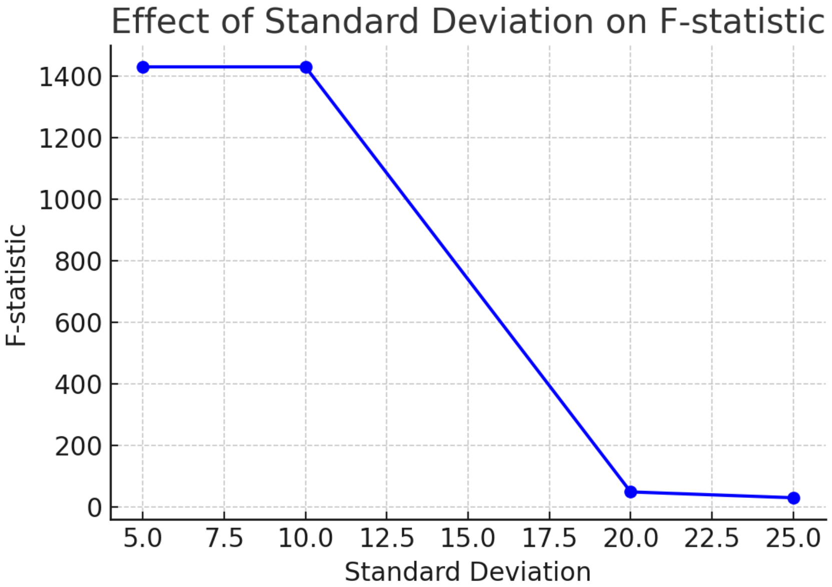


Figure 6: Effect of Standard Deviation on F-statistic

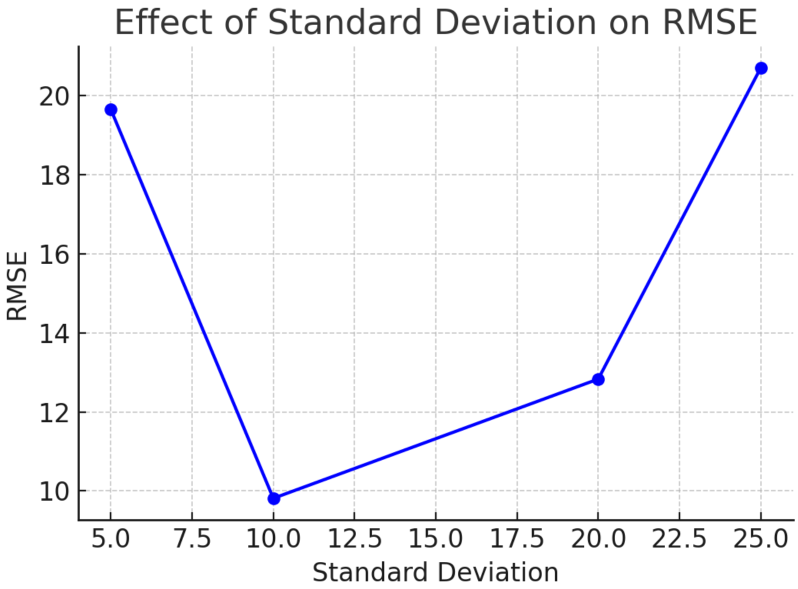


Figure 7: Effect of Standard Deviation on RMSE

Part c:

With increasing standard deviation:

- The width of confidence intervals for regression coefficients increases, indicating higher uncertainty.
- Larger confidence intervals suggest less precise parameter estimates.
- The p-values of regression coefficients increase, reducing their statistical significance.

This suggests that as noise increases, our ability to make reliable parameter estimates goes. The model becomes less stable, requiring alternative approaches such as regularization or more complex modeling techniques.