DS203 : Exercise 2

Nirav Bhattad

Last updated August 23, 2024

Question 1

Question 1

In $E1$ you have already created a dataset (y, x). Calculate the **Pearson Correlation Coefficient** (r) for this dataset and comment on the value.

We calculate the Pearson Correlation Coefficient (r) as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{0.1}$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y respectively.

Pearson Correlation Coefficient		
r	0.971127444318593	
Coefficient of Determination		
r^2	0.943088513108763	

Figure 1: Pearson Correlation Coefficient (r)

Question 2

Question 2

Using the $E1$ dataset, fit a Regression line and generate detailed regression output **using the built-in Regression functionality of the spreadsheet.** Show the Regression output in your report. It should include the regression coefficients, their associated standard errors, p-values, confidence intervals, the F-statistic, and R^2 values. (All spreadsheets create detailed Regression output – some provide a dialog based UI to do it, while in some others appropriate function(s) have to be used.)

I fitted a regression line to the dataset (y, x) using the built-in Regression functionality of the spreadsheet. The regression output is shown in the following figure.

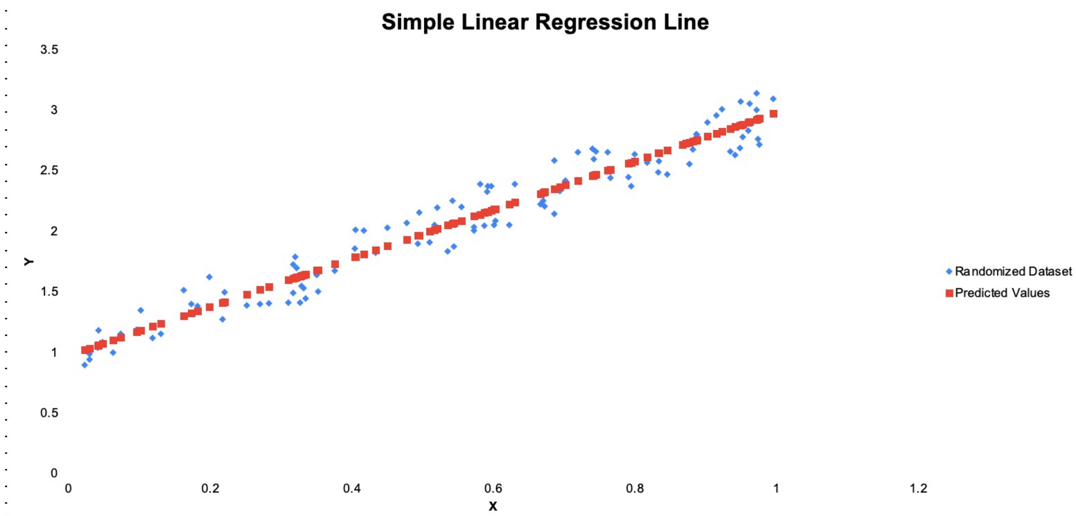

Figure 2: Regression output

This is the regression output of the dataset (y, x):

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.971127444318593					
R Square	0.943088513108762					
Adjusted R Square	0.942507783650688					
Standard Error	0.144560038370349					
Observations	100					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	33.9371296979218	33.9371296979218	1623.97223009275	8.36523045969629E-63	
Residual	98	2.0479652599764	0.0208976046936367			
Total	99	35.9850949578982				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.980513245848111	0.030976328459111	31.6536301951467	2.93641644172256E-53	0.91904173016084	1.04198476153538
X Variable 1	2.0030777006751	0.049705963538737	40.2985388084078	8.36523045969653E-63	1.90443783373749	2.1017175676127

Figure 3: Statistics Related to the Regression

These are the predicted values of y based on the regression line, also known as $\widehat{y_i}$:

RESIDUAL OUTPUT			
Observation	*Predicted Y*	*Residuals*	*Standard Residuals*
1	1.02713937735216	-0.129012604916518	-0.896991750384679
2	1.03941351519921	-0.045856517034955	-0.31882867188323
3	1.04013437363632	-0.096594495054869	-0.671596897472595
4	1.06282083328878	-0.018894361001165	-0.131367674947746
5	1.06447332797095	0.122795883308487	0.853768469989319
6	1.07706455787777	0.011917664467117	0.082860482646434
7	1.10647468384955	-0.101534179216781	-0.705941262084162
8	1.12819847851791	0.029737810875828	0.206759417402492
9	1.17287455318611	0.016827945210982	0.117000412788686
10	1.18504546445555	0.169362063435544	1.17753124842423
11	1.21845124516617	-0.094245115679125	-0.655262261644333
12	1.24207815476814	-0.086285621322078	-0.599921926642768
13	1.30687985740168	0.213492368779163	1.48435801051307
14	1.32704093987906	0.076181335916597	0.529669406293197
15	1.34528835902718	0.041840281235232	0.290904808301974
16	1.37919636261557	0.245516535357522	1.70701387621178
17	1.41657995320614	-0.139012269503215	-0.966516868854187
18	1.42161080595757	0.081417282264646	0.566073606353307
19	1.48461219606637	-0.090705076516661	-0.630649271876215
20	1.52249080455002	-0.119612624132987	-0.831636080509955
21	1.54686657477922	-0.135747274612538	-0.943816191785611
22	1.60305973319588	-0.186267613846975	-1.29507123038629
23	1.61613757805278	0.111770158221993	0.77710941445754
24	1.61615034301605	-0.121372958414079	-0.843875235971398
25	1.62240811808856	0.168974892614975	1.17483934841752

Figure 4: Predicted values of y, $\widehat{y_i}$, only 25 shown here

These are the standard errors obtained from the regression output:

SSE	2.04796525997639
MSE	0.0208976046936367

Figure 5: Standard Errors (SSE & MSE) of the regression

Question 3

Answer each of the following questions:

Question 3.1

Comment on the value of R^2 (the Coefficient of Determination). How good is the regression? What does this value represent?

The value of R^2 which I got is 0.943088513108762. The value of R^2 is a measure of how well the regression line fits the data. It is a measure of the proportion of the variance in the dependent variable that is predictable from the independent variable. The value of R^2 ranges from 0 to 1. A value of R^2 close to 1 indicates that the regression line fits the data very well. In this case, the value of R^2 is very close to 1, which indicates that the regression line fits the data very well.

Question 3.2

Independently calculate the value of R^2 using its basic definition. Compare it with the value created by the Regression functionality

We calculate the value of R^2 using the basic definition:

$$
\begin{aligned}
R^2 &= \frac{SSR}{SST} \\
&= \frac{33.9371296979218}{35.9850949578982} \\
&= 0.943088513108762
\end{aligned}
$$

We find that the value of R^2 calculated using the basic definition is the same as the value of R^2 calculated using the Regression functionality.

Question 3.3

Compare R^2 with r^2 (the square of the Correlation coefficient, calculated above. What is your observation?

The value of r^2 which I got is 0.943088513108763. The value of R^2 is the same as the value of r^2. This is because R^2 is the square of the Correlation coefficient. This also shows that all the calculations are consistent with each other.

Question 3.4

What do you understand by the Standard Error values associated with the coefficients?

The Standard Error values associated with the coefficients are a measure of the accuracy of the coefficient estimates. They give an estimate of the standard deviation of the sampling distribution of the coefficient estimates. A lower value of the Standard Error indicates that the coefficient estimate is more accurate.

The values which I got are:

- Standard Error for Regression Statisics: 0.144560038370349
- Standard Error for Intercept: 0.0309763284591115
- Standard Error for Slope: 0.049705963538737

Question 3.5

Are the coefficients of the Regression statistically significant? Justify your answer.

The coefficients of the Regression are statistically significant if the p-values associated with the coefficients are less than the significance level. The p-values which I got are:

- p-value for Intercept: $2.93641644172256 \times 10^{-53}$
- p-value for Slope: $8.36523045969653 \times 10^{-63}$

Both the p-values are less than the significance level of 0.05. Therefore, we can conclude that the coefficients of the Regression are statistically significant.

Question 3.6

What do you understand by the 95% confidence interval associated with each of the coefficients. What happens if ZERO is a part of this interval?

The 95% confidence interval associated with each of the coefficients is an interval within which the true value of the coefficient lies with 95% confidence. If ZERO is a part of this interval, it indicates that the coefficient is not statistically significant. If ZERO is not a part of this interval, it indicates that the coefficient is statistically significant.

The 95% confidence intervals which I got are:

- 95% Confidence Interval for Intercept: $[0.91904173016084, 1.04198476153538]$
- 95% Confidence Interval for Slope: $[1.90443783373749, 2.1017175676127]$

Both the 95% confidence intervals do not contain ZERO. Therefore, we can conclude that both the coefficients are statistically significant.

Question 3.7

Comment on the F-value / F-Statistic. What does it represent? Why is it important?

The F-value / F-Statistic is a measure of the overall significance of the regression. It is the ratio of the mean square due to regression to the mean square due to error. A higher value of the F-value indicates that the regression is more significant. The F-value is important because it helps us determine whether the regression is significant or not.

The F-value which I got is 1623.97223009275. This is a very high value, which indicates that the regression is very significant.

Question 4

Create 5 variants of the $E1$ dataset – by changing the variance of the data. In each case fit a regression line using the built-in regression functionality.

Question 4.1

For each variant, note down the regression outcomes and other statistics such as R^2, p-value, F-value, SSE, MSE, variance of y (use a Table to record all these values for each variant).

For creating the 5 variants of the $E1$ dataset with increasing variance, I have created the following datasets:

- Variant 1: $E1$ dataset
- Variant 2: $E1$ dataset with error term multiplied by 1.25
- Variant 3: $E1$ dataset with error term multiplied by 1.5
- Variant 4: $E1$ dataset with error term multiplied by 1.75
- Variant 5: $E1$ dataset with error term multiplied by 2

All the error terms were randomly generated using the `RAND()` function in Excel. The regression outcomes for each variant are tabulated in Figure 6.

	Variant 1		Variant 2		Variant 3		Variant 4		Variant 5	
	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope	Intercept	Slope
Coefficient	0.9991037664	2.042549164	1.03086457	1.806045268	0.9131324751	2.114564563	0.918796428	2.201071243	0.8842847786	2.120514233
Standard Deviation	0.05796753268	0.09301722344	0.06924853378	0.1111192083	0.0885792194	0.142138067	0.1130946126	0.1814765329	0.1217635932	0.1953871561
T-Stat	17.23557516	21.9588275	14.88644616	16.25322296	10.30865344	14.87683495	8.124139663	12.12868247	7.262308506	10.85288448
P-Value	0	0	0	0	0	0	0	0	0	0
Lower 95%	0.8840690844	1.857959511	0.8934431082	1.585532816	0.737349897	1.832496195	0.69436385	1.840936969	0.6426488908	1.732774781
Upper 95%	1.114138448	2.227138816	1.168286031	2.026557721	1.088915053	2.396632931	1.143229006	2.561205516	1.125920666	2.508253686
Multiple R	0.9116412924		0.8540531549		0.8325251003		0.7747072057		0.7388128699	
R Square	0.831089846		0.7294067915		0.6930980426		0.6001712546		0.5458444567	
Adjusted R Square	0.829366273		0.7266456363		0.68996639		0.5960913694		0.5412102165	
Standard Error	0.2705223364		0.3231684063		0.4133806682		0.5277888747		0.5682451917	
Observations	100		100		100		100		100	
Variance of y	0.4288855185		0.3820601851		0.5511775862		0.6896636396		0.7038138369	
F-value	482.190105		264.1672566		221.3202182		147.1049384		117.7851015	
SSE	7.171868781		10.23490624		16.74659053		27.29898743		31.64445459	
MSE	0.0731823345		0.1044378188		0.1708835768		0.2785610962		0.3229025979	

Figure 6: Regression outcomes for each variant tabulated together

Question 4.2

Create a plot of R^2 v/s variance (of y). What do you observe? Why?

The plot of R^2 v/s variance (of y) is shown in Figure 7.

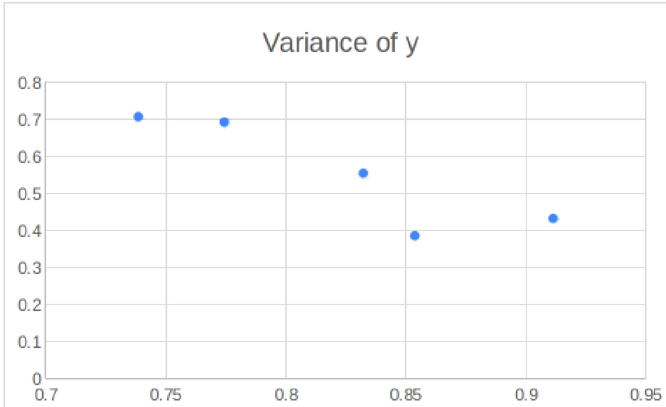

Figure 7: Plot of R^2 v/s variance (of y)

From the plot, we can observe that as the variance of y increases, the value of R^2 decreases. This is because the variance of y is a measure of how spread out the data is. When the variance of y is high, the data is more spread out and the regression line fits the data less well. This results in a lower value of R^2.

Question 4.3

Analyze the effect of variance on the regression parameters and prediction errors and state your observations and conclusions.

As the variance of y increases, the regression parameters and prediction errors are affected as follows:

- The value of R^2 decreases as the variance of y increases. This indicates that the regression line fits the data less well when the data is more spread out.
- The p-value increase as the variance of y increases. This indicates that the regression line is less statistically significant when the data is more spread out.
- The F-value decreases as the variance of y increases. This indicates that the regression line is less significant when the data is more spread out.
- The sum of squared errors (SSE) and mean squared error (MSE) increase as the variance of y increases. This indicates that the prediction errors are higher when the data is more spread out.
- The standard error values associated with the coefficients increase as the variance of y increases. This indicates that the coefficient estimates are less accurate when the data is more spread out.

In conclusion, the variance of y has a significant effect on the regression parameters and prediction errors. When the variance of y is high, the regression line fits the data less well, the prediction errors are higher, and the coefficient estimates are less accurate.