Kankanamge Harsha

PROJECT 2

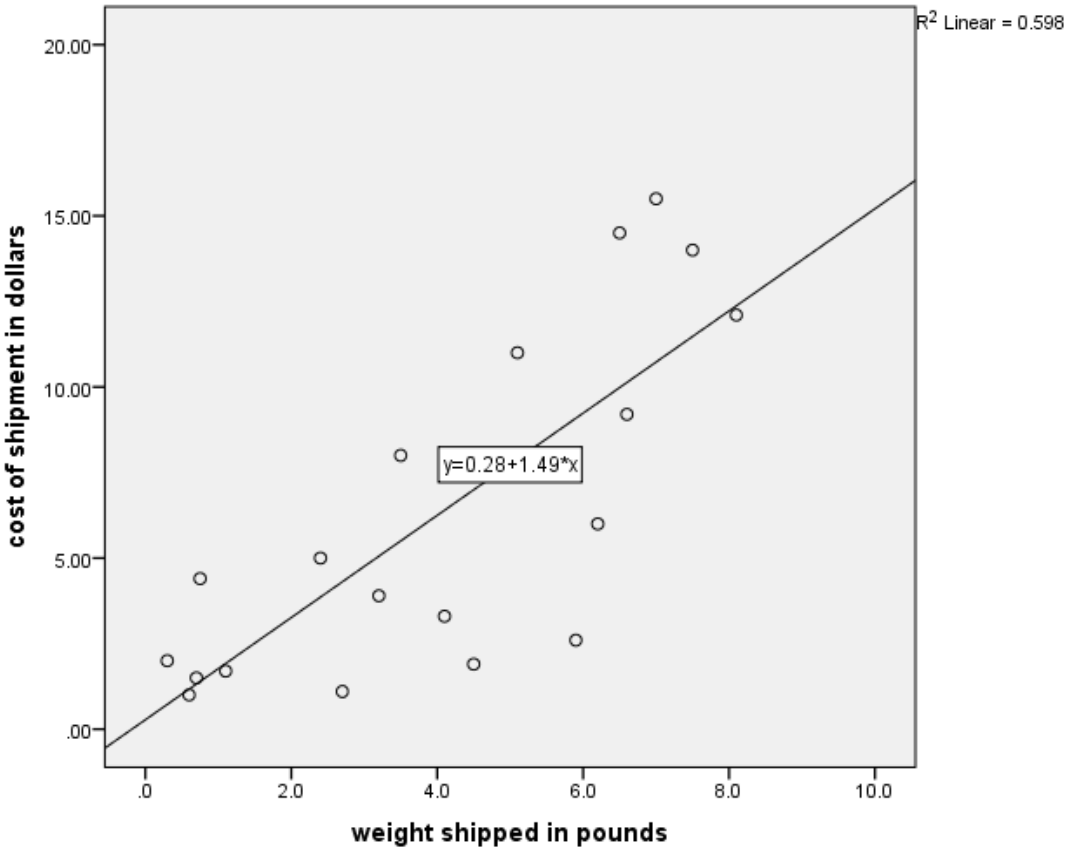
The data set **PROJ2-EXPRESS.SAV** represents the 'weight shipped' (in pounds), the 'distance shipped' (in miles) and the 'cost of shipment' (in dollars) of 20 packages received for shipment.

1)Identify the response and regressor variables in the data set.

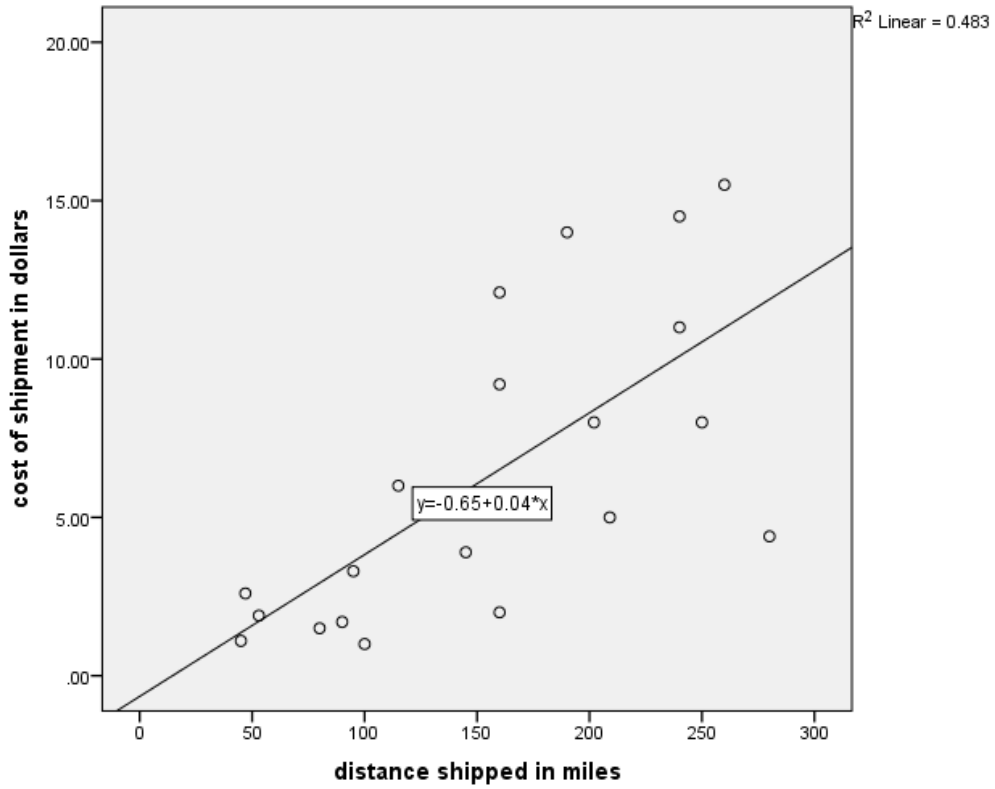
Response variable: cost (cost of shipment in dollars)

Regressor variables: weight ('weight shipped in pounds), distance (distance shipped in miles)

2) Use scatter plots to plot the dependent versus the independent variables. Interpret the plots.



Scatterplot shows strong positive linear association between “weight shipped in pounds” and “cost of shipment in dollars”. So there is a strong positive linear relationship between “weight shipped in pounds” and “cost of shipment in dollars”. Also when “weight shipped in pounds” increases “cost of shipment in dollars” also increases.



Scatterplot shows strong positive linear association between "distance shipped in miles" and "cost of shipment in dollars". So there is a strong positive linear relationship between "distance shipped in miles" and "cost of shipment in dollars". Also when "distance shipped in miles" increases "cost of shipment in dollars" also increases.

3) Find Pearson's Correlation coefficients and comment on the strength of linear association between the pairs of variables (x and y).

Correlations

		weight shipped in pounds	distance shipped in miles	cost of shipment in dollars
weight shipped in pounds	Pearson Correlation	1	.182	.774**
	Sig. (2-tailed)		.444	.000
	N	20	20	20
distance shipped in miles	Pearson Correlation	.182	1	.695**
	Sig. (2-tailed)	.444		.001
	N	20	20	20
cost of shipment in dollars	Pearson Correlation	.774**	.695**	1
	Sig. (2-tailed)	.000	.001	
	N	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Pearson Correlation is strongly significant between “weight shipped in pound” and “cost of shipment in dollars”. There is strong linear association between “weight shipped in pound” and “cost of shipment in dollars”. Because Pearson correlation is 0.774.

Pearson Correlation is strongly significant between “distance shipped in miles” and “cost of shipment in dollars”. There is strong linear association between “distance shipped in miles” and “cost of shipment in dollars”. Because Pearson correlation is 0.695.

And also p-value is very low for both of them

But “weight shipped in pound” has higher linear strength than “distance shipped in miles” with “shipment in dollars”. when compare the Pearson correlation values “weight shipped in pound” has the highest (0.774) association with “cost of shipment in dollars”.

4) Estimate the linear regression equation of WEIGHT and COST.

Linear regression equation is:

y=ß0+ß1x+Ɛ

cost of shipment in dollars^ (estimated)= b0 + b1x

b1=ß1^=1.493

cost of shipment in dollars^ (estimated) = 1.493* weight shipped in pounds + 0.276

Descriptive Statistics

	Mean	Std. Deviation	N
cost of shipment in dollars	6.3350	4.87791	20
weight shipped in pounds	4.058	2.5271	20

The data represent the cost of shipment ($) and weight shipped in pound of 20 packages with mean weight shipped 4.058 in pound and s.d 2.5271 and mean cost of shipment $6.335 s.d 4.87791.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.774[a]	.598	.576	3.17571

a. Predictors: (Constant), weight shipped in pounds

b. Dependent Variable: cost of shipment in dollars

R Square = 0.598 » 60% of variability in cost of shipment is explained by weight shipped

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	270.553	1	270.553	26.827	.000[b]
	Residual	181.533	18	10.085		
	Total	452.086	19			

a. Dependent Variable: cost of shipment in dollars

b. Predictors: (Constant), weight shipped in pounds

SSR= 270.553

SSE= 181.533

S^2=10.085

F= 26.827 F is significant. So there is an association between cost and weight.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.276	1.368		.202	.842	-2.599	3.151
	weight shipped in pounds	1.493	.288	.774	5.179	.000	.888	2.099

a. Dependent Variable: cost of shipment in dollars

t = 5.179 t also significant. So there is an association between cost and weight.

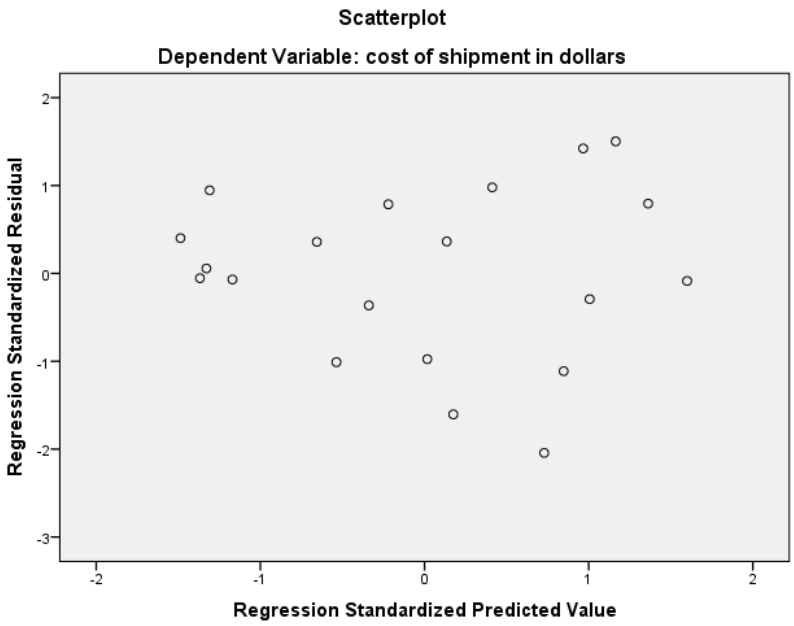
y=ß0+ß1x+Ɛ

cost of shipment in dollars^ (estimated)= b0 + b1x

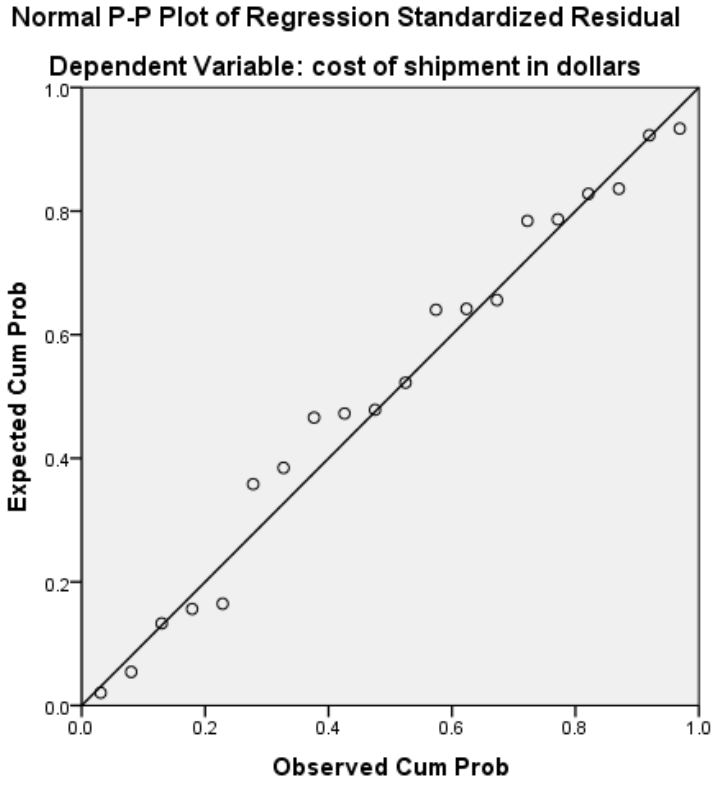
b1=ß1^=1.493

Linear regression equation:

cost of shipment in dollars^ (estimated) = 1.493* weight shipped in pounds + 0.276



Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.



Most of the data points are on or very closer to the line. We can assume residual are normal.

5)Determine and interpret the value of the slope coefficient and R^2 in this context.

cost of shipment in dollars^ (estimated) = 1.493* weight shipped in pounds + 0.276

We estimate mean "Cost of shipment in dollars" will increase by $ 1.493 for every one-pound increases in weight shipped in pound.

R Square = 0.598

Using "weight shipped in pound" the model explains 60% of the total sample variation of "cost of shipment in dollars".

6) Estimate σ^2

S^2=Estimated σ^2 = 10.085 (mean square error)

7)Estimate the multiple linear regression equation of COST using both explanatory variables. Determine and interpret the regression coefficients of WEIGHT and DISTANCE.

Descriptive Statistics

	Mean	Std. Deviation	N
cost of shipment in dollars	6.3350	4.87791	20
weight shipped in pounds	4.058	2.5271	20
distance shipped in miles	156.05	75.697	20

The data represent the cost of shipment ($), weight shipped in pound and distance shipped in miles of 20 packages with mean weight shipped 4.058 pound, s.d 2.5271, mean cost of shipment $6.335 s.d 4.87791 and mean distance shipped 156.05 miles s.d 75.697.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.957[a]	.916	.906	1.49314

a. Predictors: (Constant), distance shipped in miles, weight shipped in pounds

b. Dependent Variable: cost of shipment in dollars

R Square = 0.916 » 92% of variability in cost of shipment is explained by weight shipped and distance shipped variables

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	414.184	2	207.092	92.888	.000[b]
	Residual	37.901	17	2.229		
	Total	452.085	19			

a. Dependent Variable: cost of shipment in dollars

b. Predictors: (Constant), distance shipped in miles, weight shipped in pounds

SSR= 414.184

SSE= 37.901

SSR is high and SSE is low. So model is very good

S^2 =2.229

F= 92.888 F is strongly significant. So there is a strong association between cost and weight, distance.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-4.673	.891		-5.244	.000	-6.553	-2.793
	weight shipped in pounds	1.292	.138	.670	9.376	.000	1.002	1.583
	distance shipped in miles	.037	.005	.573	8.026	.000	.027	.047

a. Dependent Variable: cost of shipment in dollars

t = 9.376 t is significant. So there is a strong association between cost and weight.

t = 8.026 t is significant. So there is a strong association between cost and distance.

Multiple linear regression equation:

y=ß0+ß1x1+ß2x2+Ɛ

cost of shipment in dollars^ (estimated)= b0 + b1x1 + b1x2

cost of shipment in dollars^ (estimated) = 1.292* weight shipped in pounds + 0.037*distance shipped -4.673

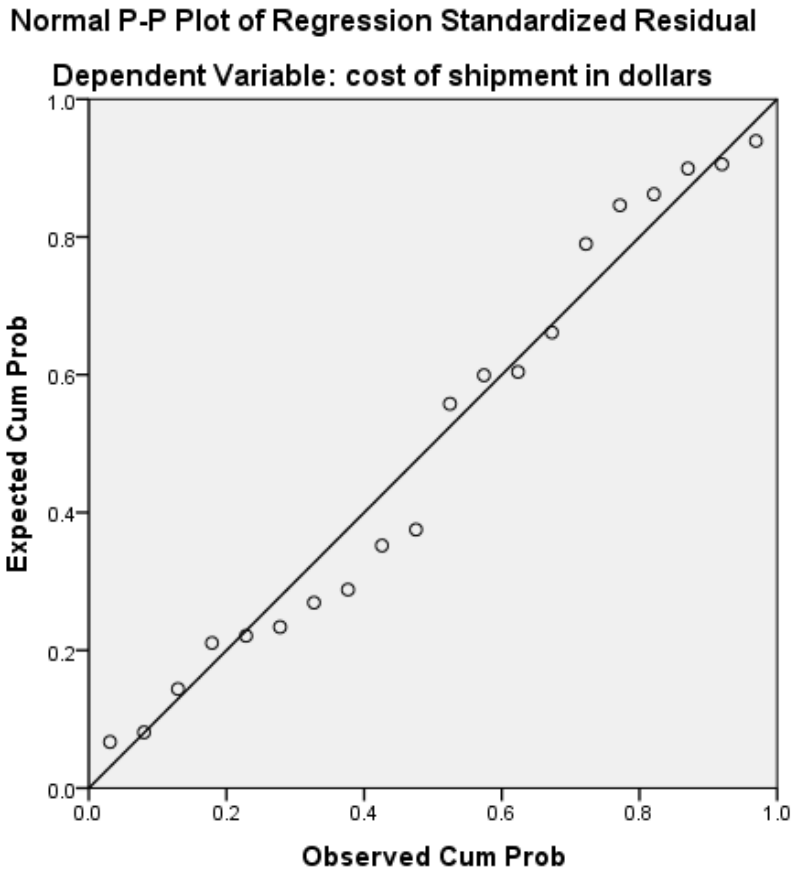
b1=ß1^= 1.292

We estimate mean "cost of shipment in dollars" will increase by $ 1.292 for every one-pound increase in the weight shipped in pound when the distance shipped variable hold constant.

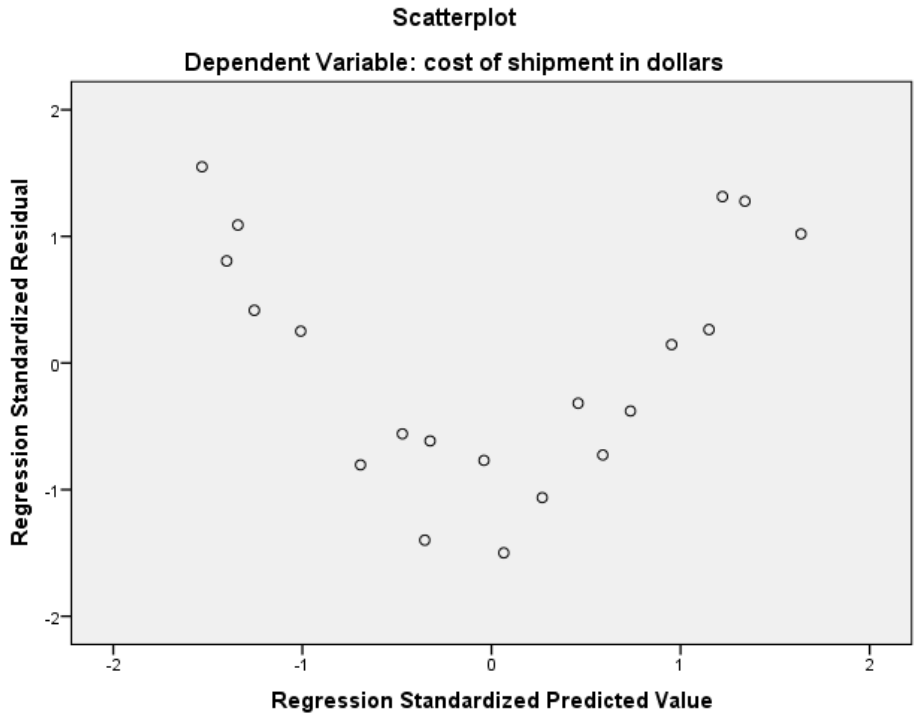
b2=ß2^= 0.037

We estimate mean cost of shipment in dollars will increase by $ 0.037 for every one-mile increase in the distance when the weight shipped variable hold constant.

b0=ß0^=0.276 does not have meaningful interpretation here.
In general, ß0^ will not have a practical interpretation under it.



Most of the data points are on or very closer to the line. We can assume residual are normal.



There is a pattern in this plot. Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

8)Test whether the regression explained by the model is significant at the 0.01 level of significance.

H0: ß1=ß2=0

H1: At least one of the line model coefficient is non zero

Test statistic F = 92.888

p-value=0.000

p-value<0.01 Reject H0

The data provide evidence that at least one of the model coefficient is non zero.
The overall model appears to be statistically good for predicting Cost of shipment.

9)Test the hypothesis that the mean COST increases as WEIGHT increases when DISTANCE is held constant. Use α = 0.01.

H0: ß1=0

H1: ß1>0

Test statistic t = 9.376 p-value for two tail test=0.000

p-value for one tail test=0.000/2= 0.000

p-value<0.01 Reject H0

We can conclude that mean COST increases as WEIGHT increases when DISTANCE is held constant.

10)Determine the 95% CI for the regression coefficient of DISTANCE and interpret the result.

95% confidence interval for ß2(distance) is (0.027, 0.047)

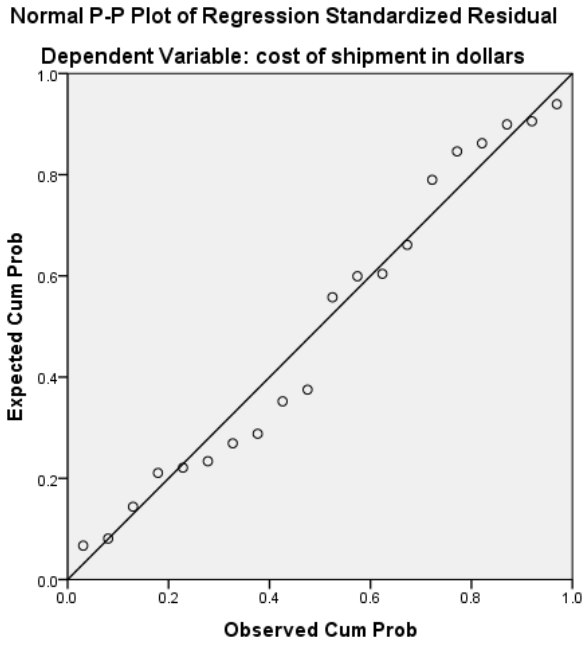
We are 95% confidence that ß2 falls between 0.027 and 0.047
With 95% confidence we can estimate mean cost of shipment increase between $0.027 and $0.047 for every 1-mile increase in shipped distances(x2) when holding the weight(x1) constant.

11)Interpret the multiple coefficient of determination.

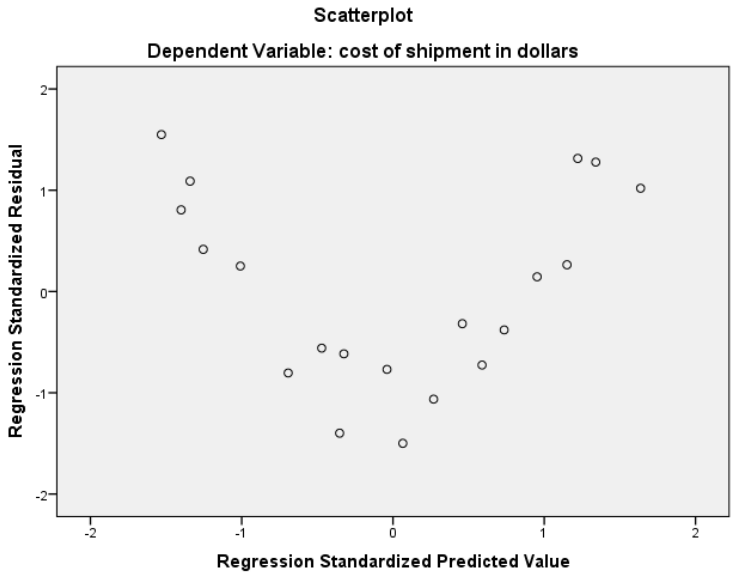
$R^2 = 0.916$

R Square = 0.916 » 92% Multiple coefficient of determination
Using “weight shipped in pound” and “distances shipped in miles” the model explains 92% of the total sample variation in cost of shipment

12)Examine the residual plots (normal probability plot and ZRESID versus ZPRED). Comment.



Most of the data points are on or very closer to the line. We can assume residual are normal.



There is a pattern in this plot. Residuals are not randomly distributed around the zero line. Therefore, variance of residuals is not homogeneous.

13)Is the model with 'WEIGHT' and 'DISTANCE' better than the model with only 'WEIGHT'? Make a table to compare the two models and explain.

Model	R^2	F	S^2	t	
				weight	distance
cost^ (estimated) = 1.493* weight+ 0.276	0.598	26.827	10.085	5.179	
cost^ (estimated) = 1.292* weight + 0.037*distance -4.673	0.916	92.888	2.229	9.376	8.026

Model with 'WEIGHT' and 'DISTANCE' better than the model with only 'WEIGHT'. Because Model with 'WEIGHT' and 'DISTANCE' has higher R^2, very higher F value(significant), small mean square error and higher significant t values when comparing model with only 'WEIGHT'. The R^2 is very high of the model with 'WEIGHT' and 'DISTANCE' (0.916 verse 0.598).

Therefore, Model with 'WEIGHT' and 'DISTANCE' is better.