CHAPTER 9

REGRESSION AND CORRELATION

1.0 Regression
2.0 Method of Least Squares
3.0 Correlation

“One day the teacher told us that four and one are five, and today she says that three and two make five. I wish that she would make up her mind.”

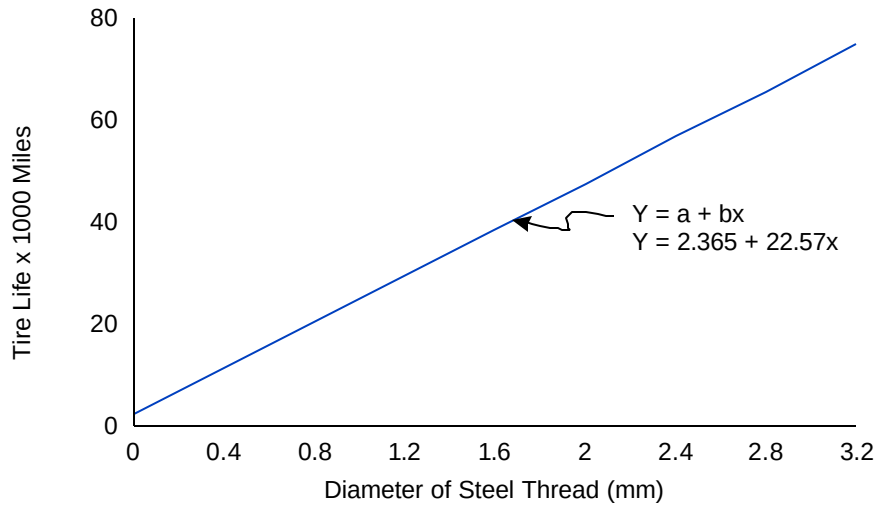
A schoolboy to his pal

REGRESSION AND CORRELATION

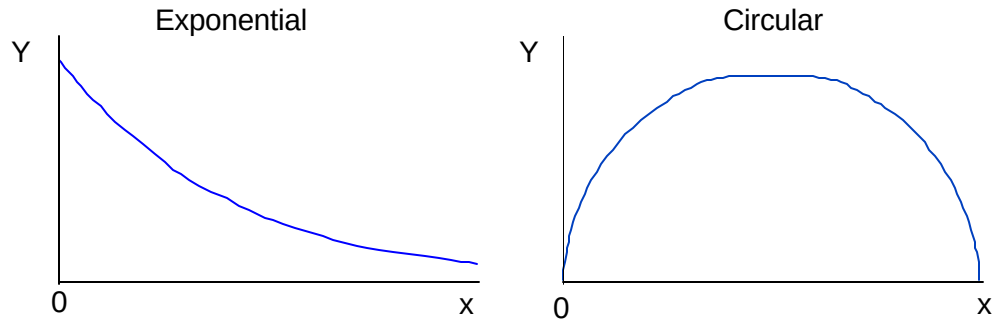
1.0 REGRESSION

Regression analysis is intended to examine the relationship of a variable Y to a variable x. The x variable is called the independent variable and the Y variable is called the dependent variable. An analysis with one independent variable is called simple regression and an analysis with more than one independent variable is called multiple regression. The levels of the independent variables are set and observations are made on the dependent variable. The objective is to find an equation to predict Y for any value of x. Regression equations may be straight lines or curves. The most widely used regression equations are simple linear models.

1.1 Simple Linear Regression Graph



1.2 Non-Linear Regression Curves



Relationships other than linear do exist, but the actual computations for them are beyond the scope of QReview. The example that follows involves simple linear regression.

2.0 METHOD OF LEAST SQUARES

One way to find the equation of a straight line, given that you have obtained some data, is called the Method of Least Squares. Many calculators have built in programs to calculate the slope and intercept that determine the equation of a straight line. The calculator is the preferred method to find these values.

General equation of a straight line: $Y = a + bx$

b = Slope,

a = Y intercept at $x = 0$

In case your calculator breaks down, the Method of Least Squares formulas to calculate the slope and intercept are:

$$\text{Slope}\,(b) = \frac{\sum(xY) - [(\sum x)(\sum Y)/n]}{\sum x^2 - [(\sum x)^2/n]}$$

$$\text{Y Intercept}\,(a) = \frac{\sum x\,[\sum(xY)] - \sum Y(\sum x^2)}{\sum x^2 - n(\sum x^2)}$$

The following data were used to find the equation of the line shown on the first page of this chapter. A sample of four is not a sufficient sample size in an actual study. Four was used in this example to simply show the calculations and methods involved. A sample size of thirty or more would be preferred.

Diameter of Thread (x)	Tire Life x 1000 Miles (Y)	xY	x^2
.2	9.0	1.80	.04
.6	11.7	7.02	.36
1.0	27.0	24.00	1.00
1.4	34.0	47.60	1.96
3.2	81.7	83.42	3.36

$b = (83.42 - 65.36)/(3.36 - 2.56) = 18.06/.8 = 22.575$

$a = (266.944 - 274.512)/(10.24 - 13.44) = -7.568/-3.2 = 2.365$

Since $Y = a + bx$, $Y = 2.365 + (22.575)x$

Now use the equation, Y = 2.365 + (22.575)x, to predict Y for any value of x.

If x = 1.2 mm, Tire Life (Y) = 2.365 + 22.575(1.2) = 29.455 K Miles or 29,455 Miles.

If x = 1.7 mm, Tire Life (Y) = 2.365 + 22.575(1.7) + 2.365 = 40.743 K Miles or 40,743 Miles.

The equation was used to plot the graph shown on page one. The graph can now be used to determine Y for any value of x.

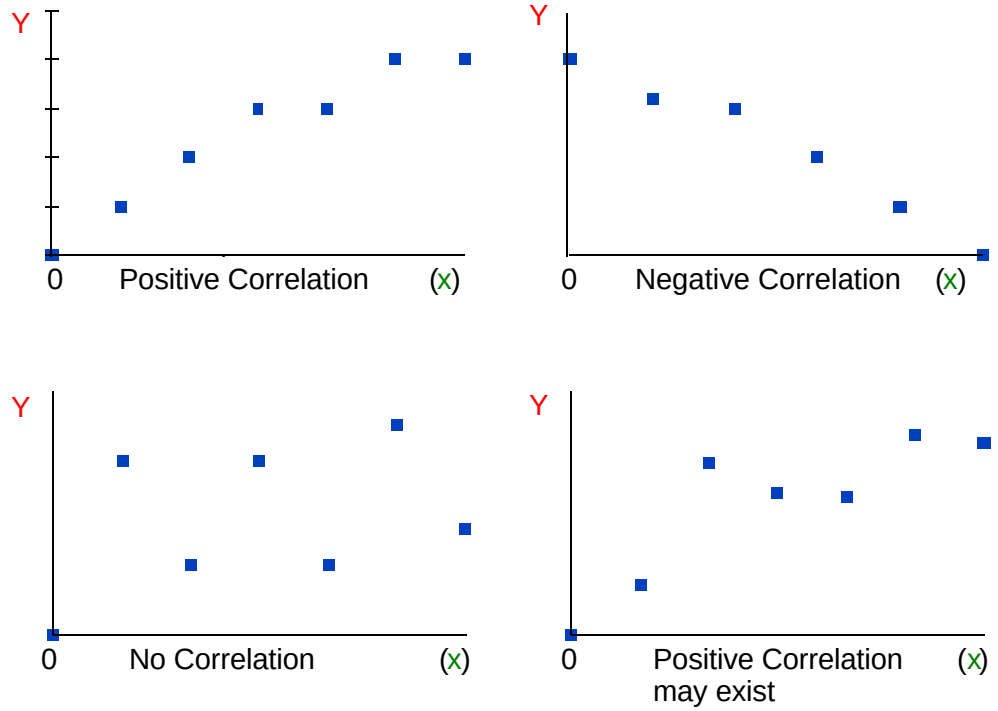
3.0 CORRELATION

3.1 General

In correlation analysis, the two variables (x and Y) are plotted to see if a relationship could exist between them. The simplest way to study correlation is to plot a scatter diagram. A single point on the diagram represents one pair of x and Y values. Statistics may be calculated to determine the strength of the relationship. The calculated statistic to determine the strength of the relationship is called the correlation coefficient and is denoted by r.

A strong correlation does not always imply that x caused Y. A good understanding of the data and where they came from is necessary to conclude that Y could logically be predicted from x even though the correlation coefficient is high.

3.2 Scatter Diagrams



The closer the points are to a straight line or to a known curve, the higher the degree of correlation. The more the points scatter, the less correlation. Perfect correlation would

exist when all points (pairs of x and Y) lie on the line or curve. Correlation also can be defined as a trend in Y with increasing or decreasing values of x.

3.3 Correlation Coefficient (r)

When a regression equation is calculated, it is important to know the degree of correlation between the two variables x and Y. The correlation coefficient is a mathematical measure of the degree of correlation.

$$\text{Correlation coefficient (r)} = \frac{\Sigma\left[(x - \bar{x})(Y - \bar{Y})\right]}{n(\sigma_x \sigma_y)}$$

The Correlation coefficient r will be a positive number if there is positive correlation between the variables, and a negative number if there is negative correlation between the variables.

Positive correlation: x increases, Y increases
Negative correlation: x increases, Y decreases

A correlation coefficient of 1, either + or -, indicates perfect correlation, but remember that perfect correlation does not imply that x and Y are related as to cause and effect. A correlation coefficient of 0 indicates no correlation.

Calculation of r from the tire life data in section 2:

(σ is calculated using n in the standard deviation formula.)

$n = 4$, $\bar{x} = .80$, $\sigma_x = .4472$

$\bar{Y} = 20.425$, $\sigma_Y = 10.4183$

$(x - \bar{x})$	$(Y - \bar{Y})$	$(x - \bar{x})(Y - \bar{Y})$
.2 - .8 = -.6	9.0 - 20.25 = -11.425	6.855
.6 - .8 = -.2	11.7 - 20.423 = -8.725	1.745
1.0 - .8 = .2	27.0 - 20.425 = 6.575	1.315
1.4 - .8 = .6	34.0 - 20.425 = 13.575	8.145
		18.060

Correlation coefficient (r) = [1/4 (18.060)]/[(.4472)(10.4183)]

r = 4.515/4.659 = .9690

A correlation coefficient of .9690 indicates a strong correlation or relationship between steel thread diameter (x) and tire life (Y).

3.4 x and Y Relationship

x and Y relationship	
r = + 1.0	Strong, positive correlation
r = + .75	Fair, positive correlation
r = + .50	Weak, positive correlation
r = 0	No correlation
r = - .5	Weak negative correlation
r = - .75	Fair negative correlation
r = - 1.0	Strong negative correlation