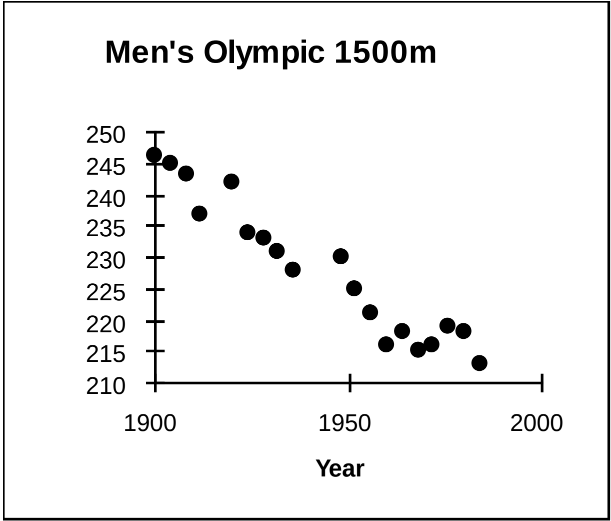
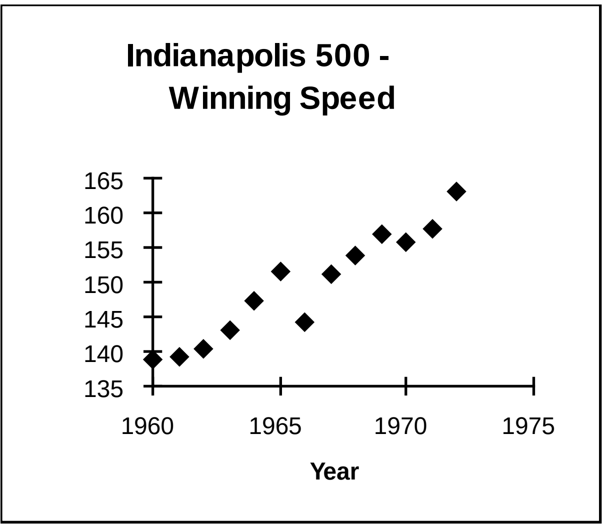
University of Stirling
Computing Science and Mathematics
Faculty of Natural Sciences

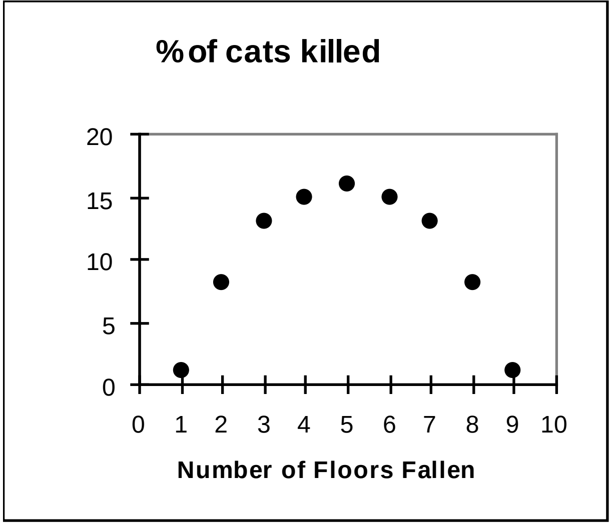
MATU9D2 : PRACTICAL STATISTICS

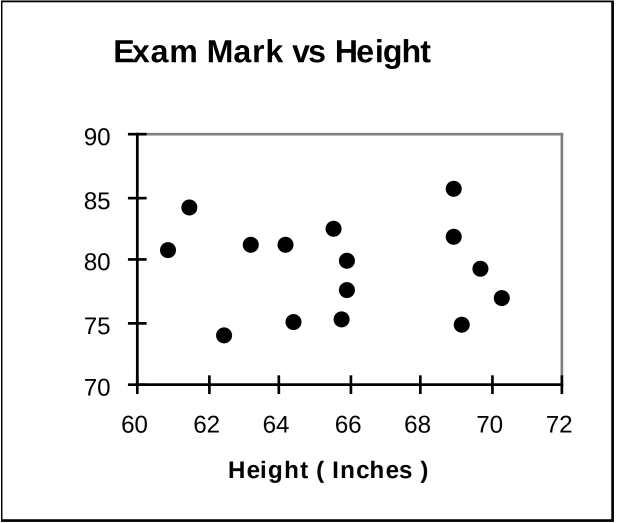
Chapter 7. Correlation and Regression

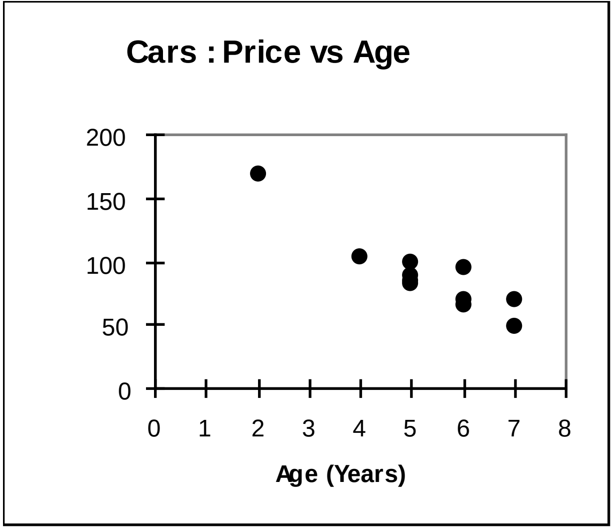
This chapter investigates different aspects of relationship between **quantitative** variables. In particular we examine **linear** relationships.

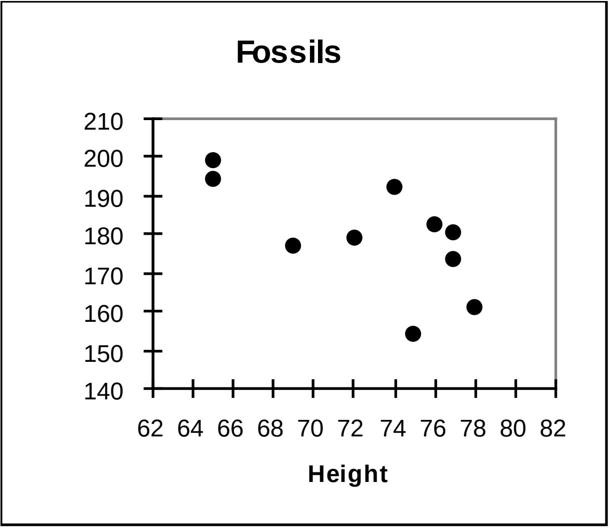

Men's Olympic 1500m


Indianapolis 500 - Winning Speed

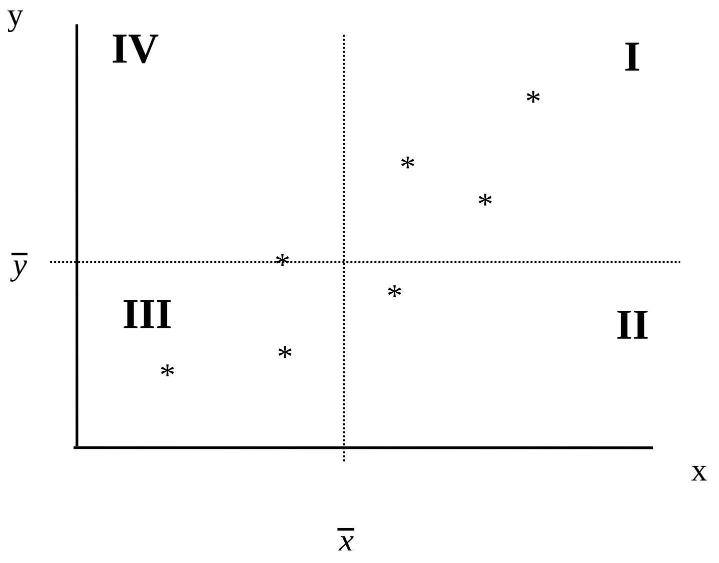

% of cats killed


Exam Mark vs Height


Cars : Price vs Age


Fossils

7.1 Covariance & Correlation



To examine the relationship, let us suppose that for every pair (x, y) we subtract $\bar{x}$ from x and $\bar{y}$ from y. Note that the mean of any set of these deviations is zero. This transformation therefore has the effect of shifting the origin of each plot to ($\bar{x}$, $\bar{y}$). The effect of this ‘transformation’ is illustrated by the secondary axes on the plot above.

Each plot is divided into four quadrants, by the new axes, they are numbered I to IV from the upper right quadrant. The points in the represent what is called ‘the joint distribution of x- $\bar{x}$ and y - $\bar{y}$. We can see that

Relationship

(i) Majority of points in quadrants I and III ....

(ii) Greater concentration of points in quadrants I and III than for a Moderate Positive Relationship ....

(iii) Majority of points in quadrants II and IV ....

(iv) Even distribution among all four quadrants ....

Since the quadrants are defined by the axes corresponding to (x- $\bar{x}$) = 0 and (y - $\bar{y}$) = 0, the x-deviation scores are positive in quadrants I and II and negative in quadrants III and IV. Similarly the y-deviation scores are positive in quadrants I and IV and negative in quadrants II and III.

Therefore, the sign of (x- $\bar{x}$)(y - $\bar{y}$) is positive in quadrants I and III
and the sign of (x- $\bar{x}$)(y - $\bar{y}$) is negative in quadrants II and IV.

Therefore, when

(I) X and Y are positively related - $\Sigma (x- \bar{x})(y - \bar{y}) > 0$ i.e. positive

(ii) The stronger the relationship the fewer negative products therefore the larger the positive value.

(iii) X and Y are negatively related - $\Sigma (x- \bar{x})(y - \bar{y}) < 0$ i.e. negative

(iv) If there is no relationship the negative deviation cancels out the positive values therefore the sum is near zero.

We therefore have a numerical value that reflects both the magnitude and direction of the statistical relationship of X and Y. However unless two sets of measurements are completely unrelated, the absolute value of the sum tends to increase with the sample size.

e.g. even if height and weight exhibit exactly the same degree of positive relationship in two different groups $\Sigma (x- \bar{x})(y - \bar{y})$ should be larger for the larger group.

For a negative relationship $\Sigma (x- \bar{x})(y - \bar{y})$ becomes smaller as n increases. We correct this by calculating the **COVARIANCE**

$$Cov(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

OR

$$Cov(X,Y) = \frac{S_{XY}}{n-1}$$

$$where \quad S_{XY} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Example Calculate the covariance for the following set of data.

	The Data		
Obs i	x_i	y_i	
1	25	48	
2	26	49	
3	24	46	
4	23	46	
5	27	47	
6	21	43	
7	24	47	
8	27	48	
9	22	43	
10	24	46	

.

The covariance has a number of important properties that make it an important statistic :

1. If X and Y are not related, Cov(X,Y) = 0

2. The magnitude of Cov(X,Y) increases as the strength of the relationship between X and Y increases

3. the sign of Cov(X,Y) corresponds to the direction of the relationship between X and Y.

Unfortunately it has one deficiency. the magnitude of Cov(X, Y) varies with the unit of measurement.

<u>Example</u> Define X as the weight to the kg, Y as height in metres and W as height in centimetres. If we take our measurements on the same n persons, the relation between height is the same irrespective of the units in which our measurements are taken, and a meaningful statistic should therefore be expected to yield the same value for weight (X) and height in m (Y) or for weight and height in cm (W).

The covariance of X and Y is $Cov(X,Y) = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$

and the covariance of X and W is $Cov(X,W) = \dfrac{\sum(x-\bar{x})(w-\bar{w})}{n-1}$

Since any height in cm is equal to 100 times height in m, w_i = 100 y_i. Then $\bar{w} = 100\bar{y}$ and the covariance of X and W can be written as

$$\frac{1}{n-1}\sum(x_i-\bar{x})(100y_i-100\bar{y}) = \frac{1}{n-1}\sum(x_i-\bar{x})100(y_i-\bar{y})$$

$$= \frac{100}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y}) = 100Cov(X,Y)$$

i.e. the covariance of weight and height measured in cm is 100 times the covariance measured in m !!!

This substantially reduces the usefulness of the covariance.

It might be useful when comparing covariances of the same measurements (height in cm and weight in m) in two different groups (e.g. two groups - men and women). However comparisons amongst covariances in different measurements would yield little information.

We can eliminate this by standardising the measurements :

$$Let \quad z_i = \frac{x_i - \bar{x}}{s_x} \quad then$$

$$Cov(Z_X, Z_Y) \quad = \quad \frac{1}{n-1}\sum (z_x - \bar{z}_x)(z_y - \bar{z}_y) \quad = \quad \frac{1}{n-1}\sum z_x z_y$$

$$= \quad \frac{1}{n-1}\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \quad = \quad \frac{1}{n-1}\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

which is defined as the **CORRELATION COEFFICIENT**, denoted by **r**.

This can be re-written as

$$\boxed{\begin{aligned} r \quad &= \quad \frac{Cov(X,Y)}{s_x \; s_y} \\ \\ where \quad s_x \quad &= \quad \sqrt{\frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]} \\ and \quad s_y \quad &= \quad \sqrt{\frac{1}{n-1}\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]} \end{aligned}}$$

OR

$$\boxed{\begin{aligned} r \quad &= \quad \frac{S_{XY}}{\sqrt{S_{XX} \; S_{YY}}} \\ where \quad S_{XX} \quad &= \quad \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \\ S_{XY} \quad &= \quad \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \\ S_{YY} \quad &= \quad \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} \end{aligned}}$$

This is also known as **<u>Pearson's Product Moment Correlation</u>**

7.2 Interpreting the Correlation Coefficient, r

When two measurements X and Y are unrelated, the correlation coefficient must, like the covariance, equal zero. At the other extreme, we can think of a situation in which X and Y are perfectly related, that is, where any value of x_i is associated with one and only one value of y. If this relationship is such that the scatter plot of X and Y is a straight line, the correlation of X and Y is 1 when the relationship is positive and -1 when the relationship is negative.

Whatever the units in which X and Y are measured, then a positive r value indicates that X and Y are positively related, and the closer the value is to +1, the stronger the relationship. A negative r value indicates a negative statistical relationship and the close the value to -1 the stronger the relationship.

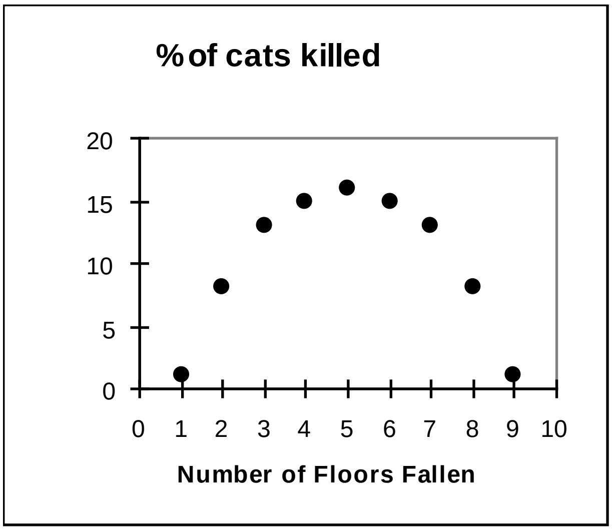
The correlation coefficient is a useful and versatile statistic, but one must be careful not to confuse correlation with causation.

For example, when reporting a high correlation between smoking and lung cancer, one is tempted to suspect that somehow smoking causes physiological changes in lung tissue. This is probably not true, but in general, a high correlation between X and Y (i.e. a value of r near 1) does not necessarily imply that X causes Y or that Y causes X.

It may simply mean that some factor W, or some combination of factors influence both. For example, children's shoe size (X) is highly correlated with spelling ability (Y). This is because shoe size and spelling are both correlated with age (W); older children have bigger feet than younger children and older children spell better.

We must also be cautious in interpreting weak correlations. A value near zero may mean that X and Y are not statistically related. However, even a perfect statistical relation will yield $r = \pm 1$ only if the relationship is linear.

A statistical relationship may be also be curvilinear. For example, cats that fall from buildings apparently reach terminal velocity after about four or five floors of fall and then if they have time extend their limbs spread-eagled which increases air resistance and slows them down. As a result the proportion of cats killed by falls from 9 floors is the same as for falls of 1 floor and is much higher for five floors. The relationship of proportion of fatalities and distance fallen is very strong but that relationship is curvilinear and would yield a correlation near zero.



‘How cats survive falls from New York skyscrapers’. J Diamond, Natural History (1989)

So the **correlation coefficient** is an index that describes the **direction and strength of a linear relationship between two measurements X and Y.**

The population correlation of the variables X and Y is denoted by ρ. It is equal to the expected value of the product of standardised random variables Z_X and Z_Y
i.e.

$$\rho_{XY} \quad = \quad E(\, Z_X \, Z_Y \,)$$

Like r_{XY}, the population correlation coefficient ρ_{XY} can assume values from -1 to 1.

7.3 Hypothesis Tests

As well as calculating ‘the r value’ we will often want to decide whether the relationship is statistically significant. The concepts are the same as for any hypothesis test i.e.

Let ρ be the population correlation then the hypotheses are

$$H_o : \rho = 0$$

and $H_1 : \rho \neq 0$

or $H_1 : \rho < 0$

or $H_1 : \rho > 0$

There are two possible statistical tests (i) for small samples n ≤ 10
and (ii) for larger samples n > 10.

(i) **<u>Small Samples n ≤ 10</u>**

The Test Statistic is $T = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

which is distributed as Student’s t with (n-2) degrees of freedom when the null hypothesis is true and if n > 4 this can be used to test the hypotheses above.

(ii) **Large Samples - n > 10**

In testing hypotheses about ρ, the joint distribution of X and Y is assumed to be bivariate normal. This implies three properties (i) the marginal distributions of X and Y are normal (ii) all conditional distributions of Y on x and X on Y are normal and (iii) X and Y are independent if and only if ρ_{XY} is zero.

The property that $\rho_{XY} = 0$ implies independence of X and Y is the basis for both the following test and the small sample test introduced previously.

The Formal Test is based on Fisher's Z_r

If the joint distribution of X and Y is bivariate Normal and the number of pairs x_i, y_i is greater than 10, the most useful test statistic is based on Fisher's transformation of the correlation to a normally distributed random variable which we denote by Z_r.

$$Z_r \quad = \quad \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

where ln denotes the natural logarithm of the value in the brackets. For moderate sample sizes (i.e. n > 10), the distribution of Fisher's Z_r is approximately normal with

$$E(Z_r) \quad = \quad \frac{1}{2}\ln\left(\frac{1+\rho_{XY}}{1-\rho_{XY}}\right) \qquad and \qquad V(Z_r) \quad = \quad \frac{1}{n-3}$$

i.e. If ρ is the correlation of a bivariate normally distributed random variable, then for n > 10 the Test Statistic is

$$\frac{Z_r}{\sqrt{1/(n-3)}}$$

which is distributed approximately N(0,1) and can be used to test the hypotheses that ρ = 0.

Example

Many scientists who study animal behaviour are interested in the relationship between social dominance and reproductive success. The following data are wins in aggressive encounters, number of cubs born (1978-1982) and number of cubs surviving 1 year for 12 female spotted hyenas observed in the Masai Mara National Reserve in Kenya.

Female	Wins	Cubs Born	Cub Survival
04	63	5	5
03	45	6	6
63	11	2	no data
N2	10	5	1
KB	3	4	2
40	9	5	2
30	4	5	3
22	3	3	2
11	3	1	no data
44	5	2	no data
16	2	3	3
31	3	3	1

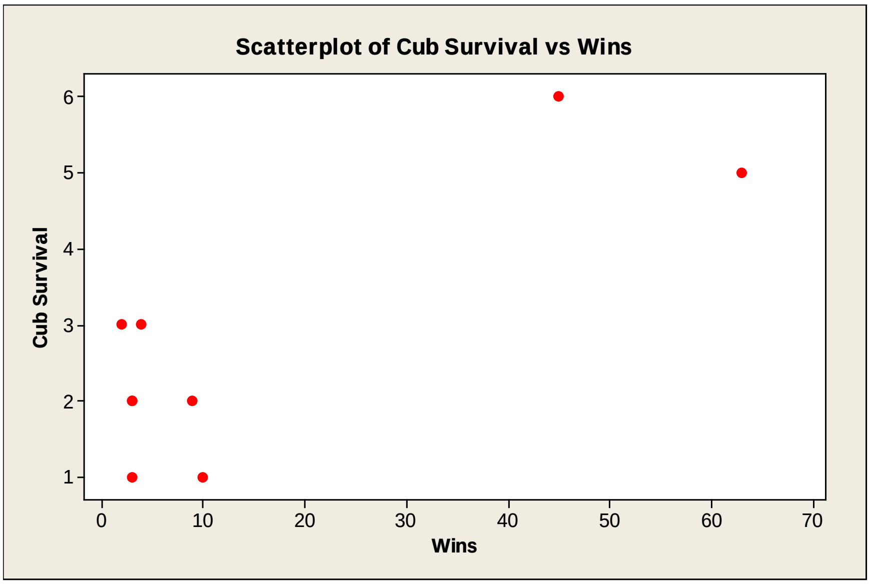
Questions

(i) Is there a significant correlation between the number of dominance encounters won and the number of cubs surviving 1 year?

(ii) Is the correlation between wins and cubs born significantly different from zero?

Solutions

(i) Let X = number of encounters Y = number of cubs surviving 1 year



Hypotheses

Significance Level

Test Statistic

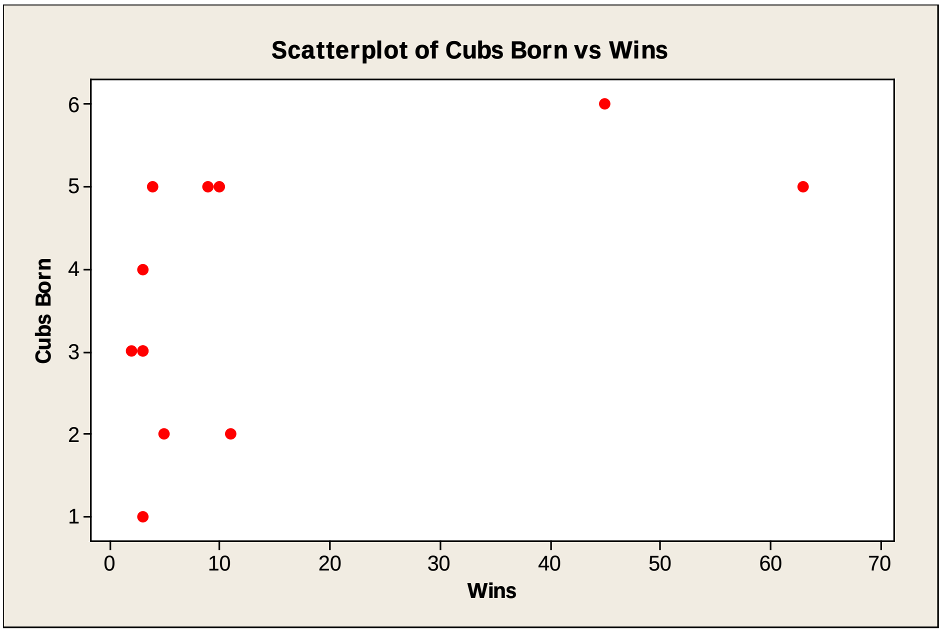
Observed Test Statistic

Rejection Region

Conclusion

(ii) Is the correlation between wins and cubs born significantly different from zero?

Let X = number of encounters Y = number of cubs born



Hypotheses

Significance Level

Test Statistic

Observed Test Statistic

Rejection Region

Conclusion

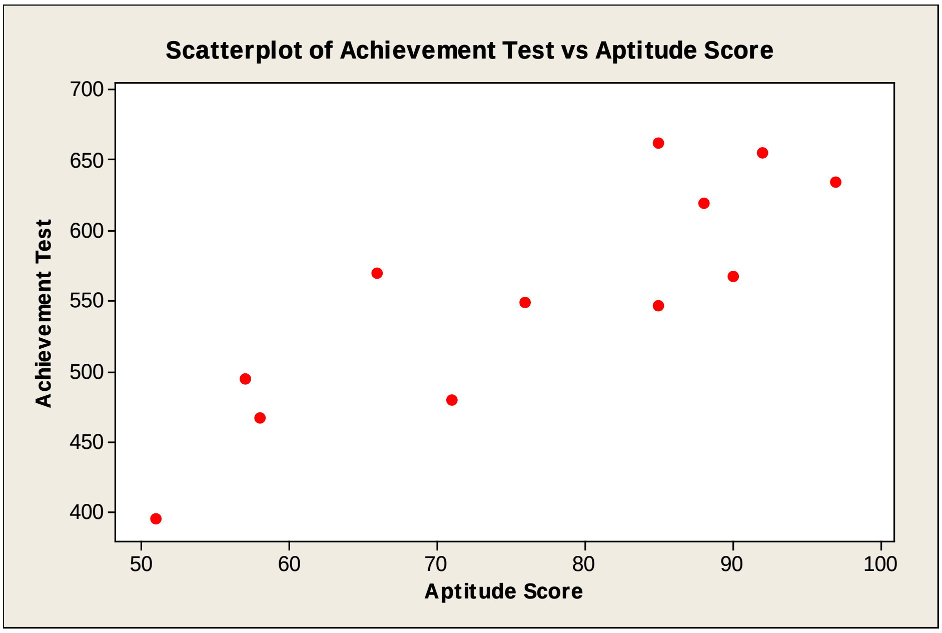
Example The following data examines the relationship between a Mathematics Achievement Test Score (y) and a Quantitative Aptitude Score (x) for a random sample of 12 students.

Student	1	2	3	4	5	6	7	8	9	10	11	12
x	88	57	76	97	71	90	66	58	92	85	51	85
y	620	495	549	635	480	568	570	467	655	547	395	662

The following graph shows the relationship. When the r value is calculated it is approximately equal to 0.865.

Does this value of r agree with what we can see in the graph?

Solution



7.4 R^2, the Coefficient of Determination - Another Measure of the Strength of the Linear Relationship Between X and Y

The sign of r tells us whether y increases or decreases as x increases, but its numerical value is difficult to interpret i.e. r = 0.5 is not halfway between a perfect correlation and no correlation.

R^2 - the Coefficient of Determination overcomes this problem and provides useful information.

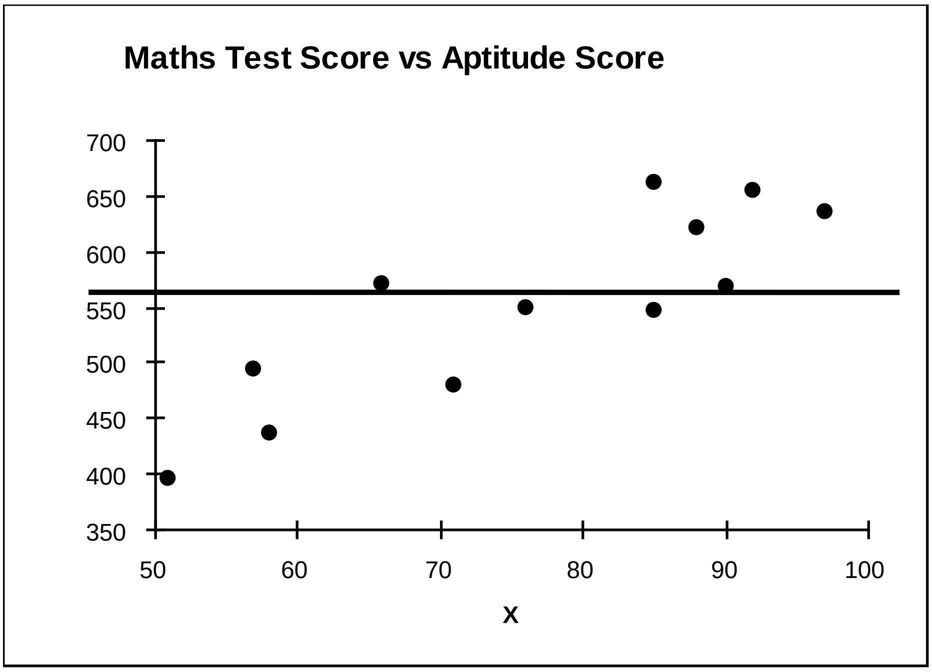
A Practical Interpretation of R^2

The best way to determine whether x and y are related is to see whether x allows us to predict y with a smaller error of prediction than we would have if we did not use x at all. If we did not use x to predict y, our best predictor of some new y value would be the sample mean.

Example The best predictor of a new student's Mathematics Achievement Test Score, based only on the Achievement Scores, would be the mean of the sample of 12 scores i.e. $\bar{y} = 551.08$

The variation of the y-values about their predicted value for this "no other information" predictor would be measured by the sum of squares of deviations S_{YY} of the y-values about their mean.

This is illustrated on the graph below :

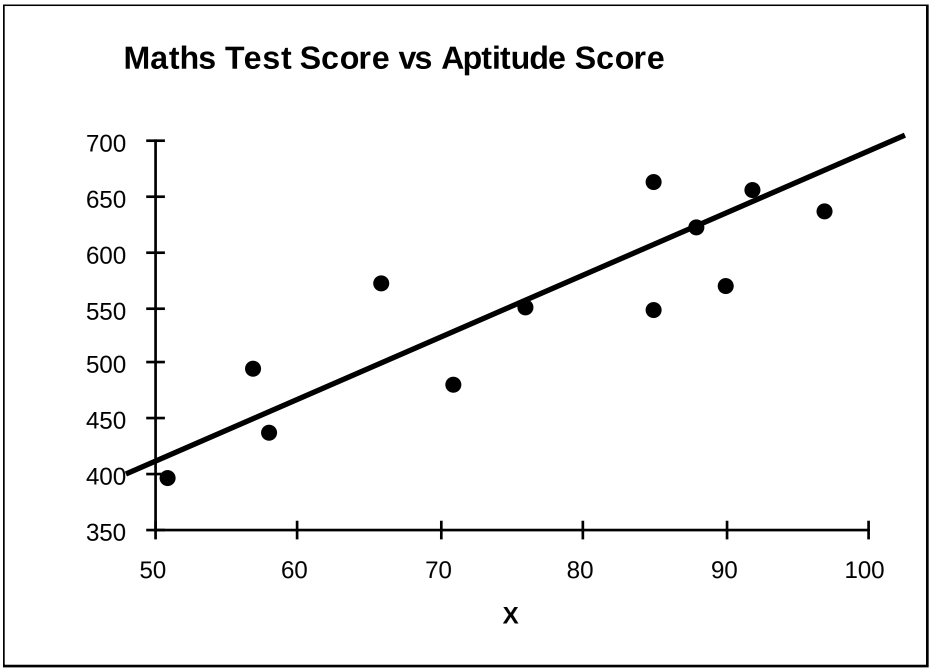


Since S_{YY} measures the variation in the prediction errors when no x variable is used to aid in prediction, S_{YY} is called the **total sum of squares of deviations**

However, if we include an x variable to aid in the prediction we get....

The technique used to use the x variable to help predict y is called least squares regression. We will be looking at this technique in detail shortly, however, we will illustrate what happens now ....

We fit a least squares line to the data e.g. the Achievement - Aptitude data. The sum of squares of deviations of the observed values of y about their predicted values (i.e. the values predicted by the least squares line) is reduced to RSS, **Residual Sum of Squares**. This is illustrated on the graph below :



Hopefully, you can see from the graphs that the deviations of the points from the least squares line, measured by RSS, are much smaller than the deviations of the points from $\bar{y}$, measured by S_{YY}.

The reduction in the sum of squares of the errors in prediction explained by the variable x is

$$S_{YY} \quad - \quad RSS \quad = \quad Total\ Sum\ of\ Squares \quad - \quad \mathrm{Re}sidual\ Sum\ of\ Squares$$

This reduction, expressed as a proportion of the Total Sum of Squares is equal to the **Coefficient of Determination, R^2.**

Therefore : R^2 is the proportion of the total variability of the y-values that can be explained by the x variable.

$$R^2 \quad = \quad \frac{S_{YY} \quad - \quad RSS}{S_{YY}}$$

Example The Mathematics Achievement Score and Aptitude Test Score

The coefficient of determination for this data is equal to

....

What does this mean?

Solution

7.5 Linear Regression

Example A lecturer in interested in whether students who do well in the Mid-Semester Test do well in the End of Semester Examination and whether students who do poorly in the Test do poorly in the Examination.

In this case the only aspect of interest is the degree to which the two methods of assessing student performance are in agreement. In some situation the lecturer might actually want to predict one score from the other. e.g. if a student is ill for the final examination - can we predict his likely grade.

In this case we will want to 'fit a model to the data'. The simplest model is a linear model - i.e. linear regression. We express the 'perfect' linear relationship between two variables by a straight line.

$$y = \alpha + \beta x$$

This is the simplest of all equations and linear predictions are simple to calculate and easy to understand.

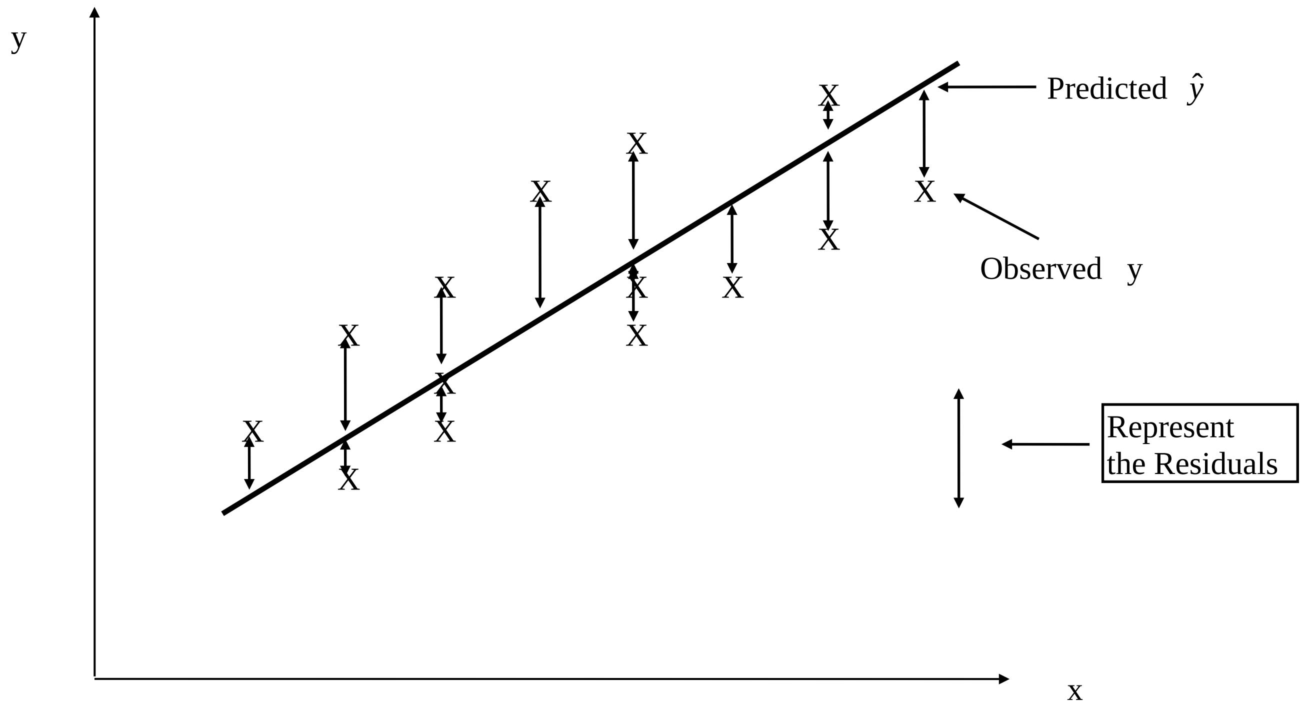
Seldom does a scatter plot of real data look anything like a line but a scatter plot showing a linear relationship will look like an ellipse with many different values of y for each value of x. We want to find the linear equation that 'best fits' a scattering of data points.

The above equation is determined once we know the values of α and β (the intercept and slope respectively), so the task is to find the values of α and β that give the best summary. The equation that best summarises the scatter plot of (X,Y) can then be used to predict or estimate the y-value associated with a specified x-value.

To understand what we mean by 'best' summary, let us suppose that the data point (x_i, y_i) represents the scores obtained by individual i and that $\hat{y}_i$ is the y-value predicted for individual i by the equation

$$\hat{y} = \alpha + \beta x$$

In general, you expect the prediction to be wrong i.e. we expect that $y_i \neq \hat{y}_i$ because we know that any line will miss most data points. The difference $(y_i - \hat{y}_i)$ is the error of prediction. This is illustrated below.



The line that best fits a joint distribution of observations is taken to be the equation that yields the smallest overall error of estimate. We know that a good statistic should consider all observed values so we might try to capture the associated with our prediction equation by adding all of the prediction errors. This tactic will be confounded by the fact that some positive errors will cancel some negative errors in an uncontrollable way.

To avoid this problem we square every error $(y_i - \hat{y}_i)$ and take the overall error of prediction to be the mean of the squared errors for all n individuals in the data :

$$\frac{\sum(y_i - \hat{y}_i)^2}{n}$$

We therefore want the line that yields the smallest possible average of the squared errors. This is sometimes called the variance of estimate. This method is called **least squares** and the line the least squares line since we are effectively minimising $\sum (y_i - \hat{y}_i)^2$.

The average squared error of prediction is least when

$$\boxed{\beta \;=\; r\,\frac{s_Y}{s_X} \quad \textit{and} \quad \alpha \;=\; \bar{y} - \beta\bar{x}}$$

$$\textit{where} \quad \bar{x} \;=\; \frac{\sum x}{n} \qquad \bar{y} \;=\; \frac{\sum y}{n} \qquad \textit{and} \quad r \textit{ is the correlation}$$

$$s_x = \sqrt{\frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]}$$

$$\textit{and} \quad s_y = \sqrt{\frac{1}{n-1}\left[\sum y^2 - \frac{\left(\sum y\right)^2}{n}\right]}$$

Calculation of β requires that we first calculate r, s_X and s_Y

or we can calculate β directly as below.

$$\boxed{\beta \;=\; \frac{n\sum xy - \sum x\sum y}{n\sum x^2 - \left(\sum x\right)^2}}$$

OR

$$\beta \;=\; \frac{S_{XY}}{S_{XX}}$$

$$\textit{where} \quad S_{XY} \;=\; \sum xy \;-\; \frac{\sum x\sum y}{n}$$

$$\textit{and} \quad S_{XX} \;=\; \sum x^2 \;-\; \frac{\left(\sum x\right)^2}{n}$$

Example In 1964 Kitty Genovese was repeatedly stabbed by an assailant in a New York street. The assault proceeded for over half an hour and although 38 of the victim’s neighbours witnessed the attack from their windows, not one person intervened or called the police. This raised a question that was subsequently the focus of major research in psychology : under what conditions will people intervene in an emergency? One feature that emerged was the Diffusion of Responsibility i.e. the more people present, the more reluctant people are to “become involved”.

The experiment below was set up to test the hypothesis that willingness to intervene is related to group size.

Each subject is an undergraduate waiting outside an office to keep an appointment with his Adviser of Studies. When the subject arrives, there are already 1, 2, 4, 5 or 11 other persons waiting. A minute or two after the subject arrives, the crash of a heavy bookcase falling over is heard from an adjacent room, followed by cries for help that become increasingly desperate as time passes.

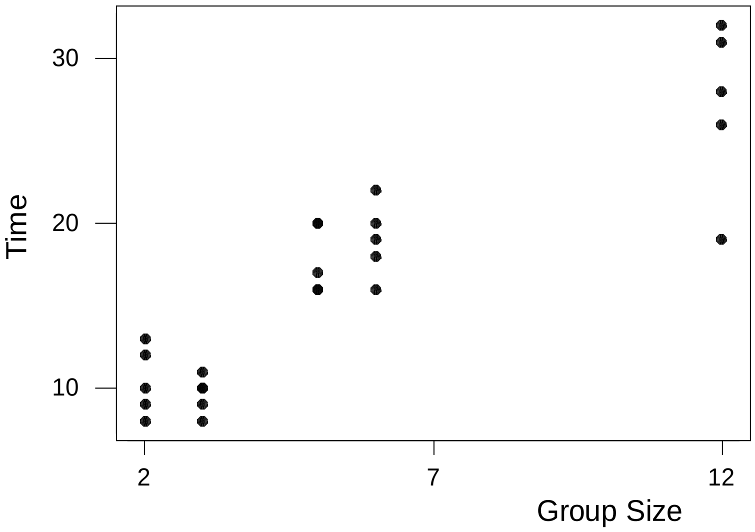
The crash and the victim’s calls for help are tape recorded and the other persons waiting are in league with the experimenter. The length of time a subject waits before investigating is Y and X is the number of persons (subject plus others) on the scene.

Data for this experiment for 25 individuals are presented below. What time would be predicted for an individual in a group of 4 persons ? Of 38 persons ?

	Group Size (x)				
	2	3	5	6	12
Time (y)	13	9	20	22	32
	9	11	20	19	26
	10	10	16	20	19
	12	10	17	18	28
	8	8	16	16	31

Comments & Solution

1. Plot the data



Independent and Dependent Variables

If we have measured two variables X and Y and found that there is a significant correlation between them, then the relationship between them can be described by a straight line. As with correlation the existence of a regression line does not imply a causal relationship.

The regression / model we have just examined is referred to as a **linear regression of Y on X**.

The model is $\boxed{y \quad = \quad \alpha \quad + \quad \beta\, x}$

where α is the intercept,
β is the slope,
X is the independent or explanatory or predictor variable,
and Y is the dependent or response variable.

It is crucial that the variables are correctly defined when performing a regression because a regression of Y on X is not the same as a regression of X on Y. They make different assumptions about the data and result in different estimates of the fitted line. (This is unlike correlation when the order of the variables in unimportant).

Explained and Unexplained Variation

For each value of X there is a predicted value from the best fit line, so

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}\, x$$

We could show that

$$\sum\left(Y_i - \bar{Y}\right)^2 = \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y}_i - \bar{Y}\right)^2$$

	Total Variation	=	Unexplained Variation	+	Explained Variation
i.e.	Total SS	=	Residual SS	+	Regression SS

The **Coefficient of Determination, R^2** is the explained variation divided by the total variation. It has a value between 0 and 1 (alternatively 0% and 100%).

Assumptions of Linear Regression

We are making three important assumptions when performing linear regression :

1. The Y values are drawn from a Normal distribution. If the distribution is significantly skewed, a ‘normalising’ transformation may be necessary.

2. The errors in X are negligible. It follows that a regression of X on Y is not the same as the regression of Y on X. This is often an unrealistic assumption!

   The correlation coefficient, r, is still valid even when both X and Y have errors.

3. The variance of Y is constant, irrespective of X. This assumption is sometimes called Homoscedasticity. This is usually a reasonable assumption but the data should always be examined for possible heteroscedasticity.

NOTE All these assumptions should be checked when interpreting the results of linear regression. See later.

7.6 Questions / Procedures associated with Linear Regression

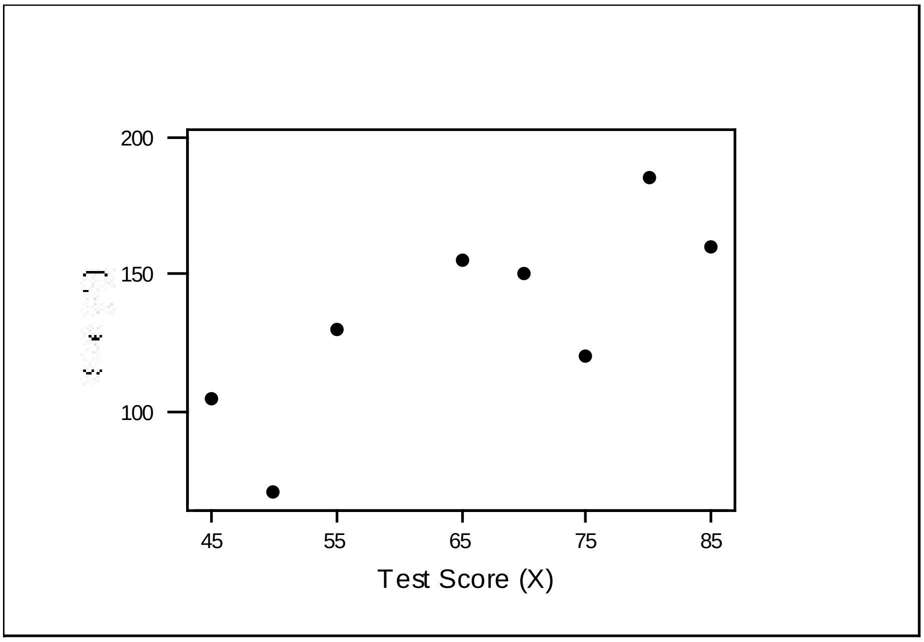
We will describe the possibilities using the following example.

Suppose that for a random sample of eight salesmen their first-year sales and test scores as trainees are as presented below. Find the fitted line which describes the relationship.

First year sales (£ 000's)	Test Score
y	x
105	45
120	75
160	85
155	65
70	50
150	70
185	80
130	55

I Plot the Data - Scatter Plot

We wish to predict the Sales (Y) from the Test Score (X) so....



II Determine the Fitted Line from the Data

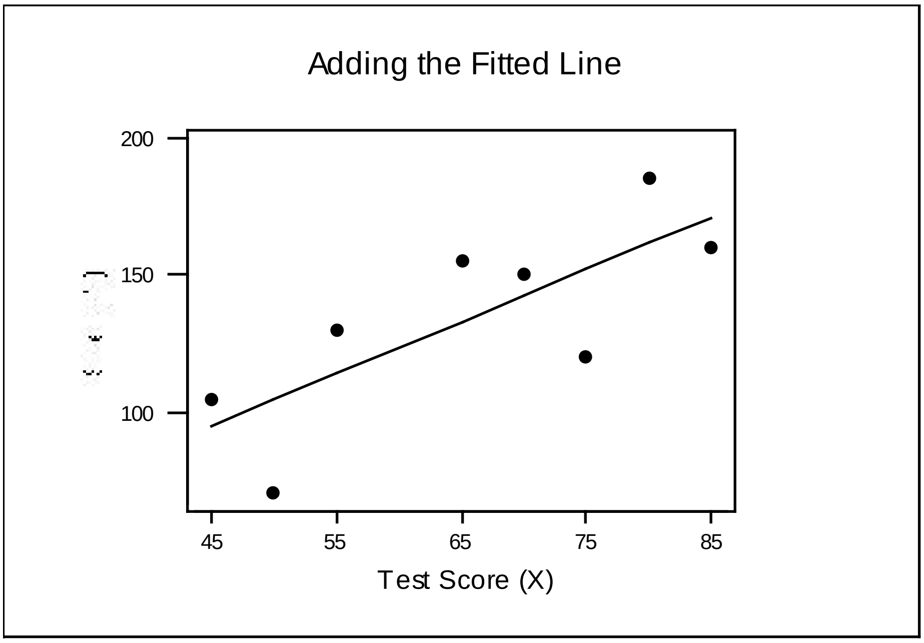
The scatter diagram above shows that there appears to be linear relationship between the variables.

III Plotting the Fitted Line on the Graph

We should now plot the fitted line on the graph i.e. plot the line y = 9.4 + 1.904 x in our example. Note that two points will determine a straight line so we could use any two values of x to determine the corresponding predicted / fitted values of y and joining these points by a straight line will produce the fitted line.

It is, however, a good idea to use the minimum and maximum values of x since the regression and thus the fitted line only applies within the range of x values in the data. i.e. we should not extrapolate much outwith the range of x.

Example



IV Predicting Values of Mean y

For a particular value of x (within the range of the sample data) the corresponding point on the line gives us the '**predicted value of y**'. For greatest accuracy instead of reading off the graph, plug the value of x into the equation for the fitted line and calculate the predicted value of y.

Example Predict the Mean First Year Sales for a Test Score of 60

For x = 60, 'predicted y' = 9.4 + (1.904 x 60) = 123.6.

This gives an estimate of the **Mean** First Year Sales of salesman with Test Scores of 60.

i.e. The mean First Year Sales for salesmen with Test Scores of 60 is £ 123, 600.

We may also calculate a Confidence Interval for the Mean Sales which we would predict for test scores of 60 (any other test score within the range of the sample data) to give an idea of the precision of the estimate of first-year sales for test scores of 60.

CONFIDENCE INTERVALS FOR PREDICTED VALUES OF MEAN Y

A 95% Confidence Interval for the predicted value of y at some value of x = x_o is

$$\left(\hat{\alpha} + \hat{\beta} x_o \right) \pm t(n-2, 0.025) \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{XX}} \right)}$$

where

$$\hat{\sigma}^2 = \frac{S_{YY} - \frac{S_{XY}^2}{S_{XX}}}{n-2}$$

Assumptions The data points are distributed approximately normally about the regression line in the y direction. The distribution is the same for all values of x.

V Hypothesis Test for the Slope of the Regression Line

For the Salesmen Example we calculated that $\hat{\beta}$ = 1.904. This is the slope of the regression line for the sample data. It is our estimate of the increase in y (first year sales) for unit increase in x (test score).

We can think of the regression line for a population of salesmen with the equation

$$y = \alpha + \beta x$$

where β is the slope and α is the intercept. So our estimate of β is the sample estimate provided by $\hat{\beta}$ = 1.904.

Could such a sample have arisen if the population value of β had been zero (implying a horizontal regression line) ?

We can perform a hypothesis test to answer this question.

Hypotheses $H_o : \beta = 0$

$H_1 : \beta \neq 0$

Significance Level : 0.05

Test Statistic :

$$T = \frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}}$$

Observed Test Statistic

Rejection Region

Conclusion

NB This test is equivalent to performing a test of the Correlation i.e.

Hypotheses $H_o : \rho = 0$
$H_1 : \rho \neq 0$

Significance Level : 0.05

Test Statistic :

Observed Test Statistic

Rejection Region

Conclusion

VI 95% Confidence Intervals for the Slope and Intercept

(a) 95% Confidence Intervals for the Slope

$$\hat{\beta} \qquad \pm \qquad t(n-2;0.025)\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$$

Example

(b) 95% Confidence Intervals for the Intercept

$$\hat{\alpha} \qquad \pm \qquad t(n-2;0.025)\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}\right)}$$

Example

VII Prediction Interval for a Single Future Observation of y for given x

95% Prediction Interval for an individual observation is

$$\hat{\alpha} + \hat{\beta}x \quad \pm t(n-2;0.025)\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{XX}}\right)}$$

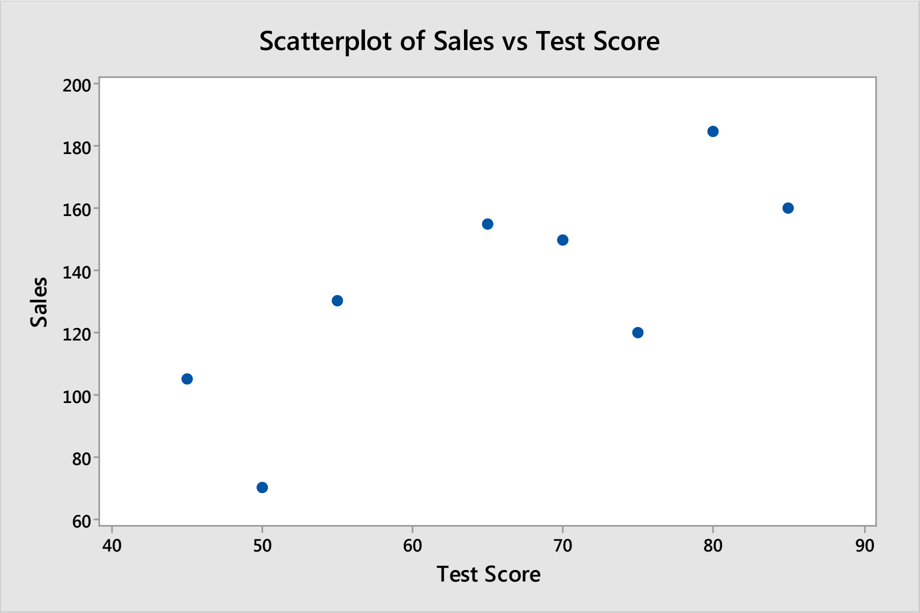
Example

7.7 Correlation & Linear Regression Using Minitab

Example 1 (Previously completed ‘by hand’ in lectures)

Sales	Test Score
105	45
120	75
160	85
155	65
70	50
150	70
185	80
130	55

Step 1 Plot the data



Step 2 Correlation

Correlations: Sales, Test Score

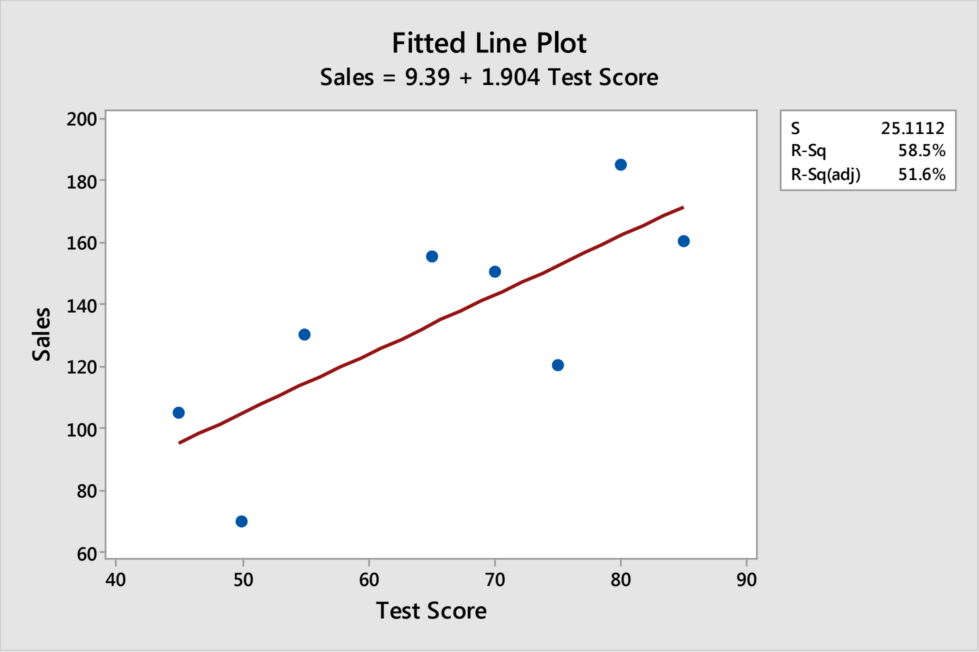
```
Pearson correlation of Sales and Test Score = 0.765
```

Step 3 Is the correlation significantly different to zero?

Correlation: Sales, Test Score

```
Pearson correlation of Sales and Test Score = 0.765
P-Value = 0.027
```

Step 4 Add the fitted line to the graph

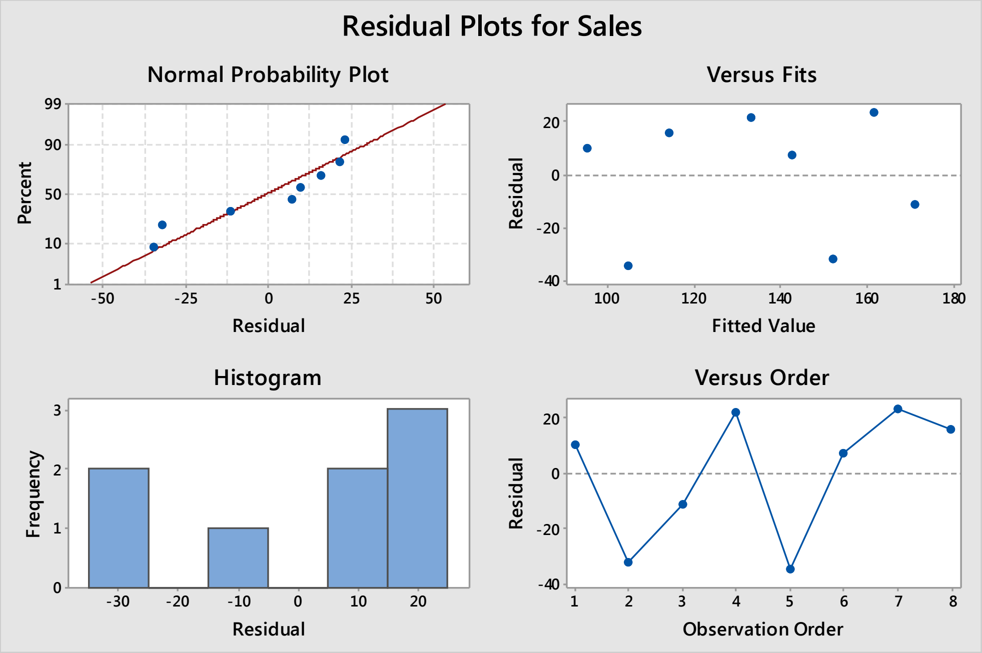


Step 5 Perform Linear Regression

Regression Analysis: Sales versus Test Score

```
Analysis of Variance

Source         DF  Adj SS  Adj MS  F-Value  P-Value
Regression      1    5338  5338.4     8.47    0.027
  Test Score    1    5338  5338.4     8.47    0.027
Error           6    3783   630.6
Total           7    9122


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
25.1112  58.52%     51.61%      27.08%


Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant      9.4     43.9     0.21    0.837
Test Score  1.904    0.655     2.91    0.027  1.00


Regression Equation

Sales = 9.4 + 1.904 Test Score
```

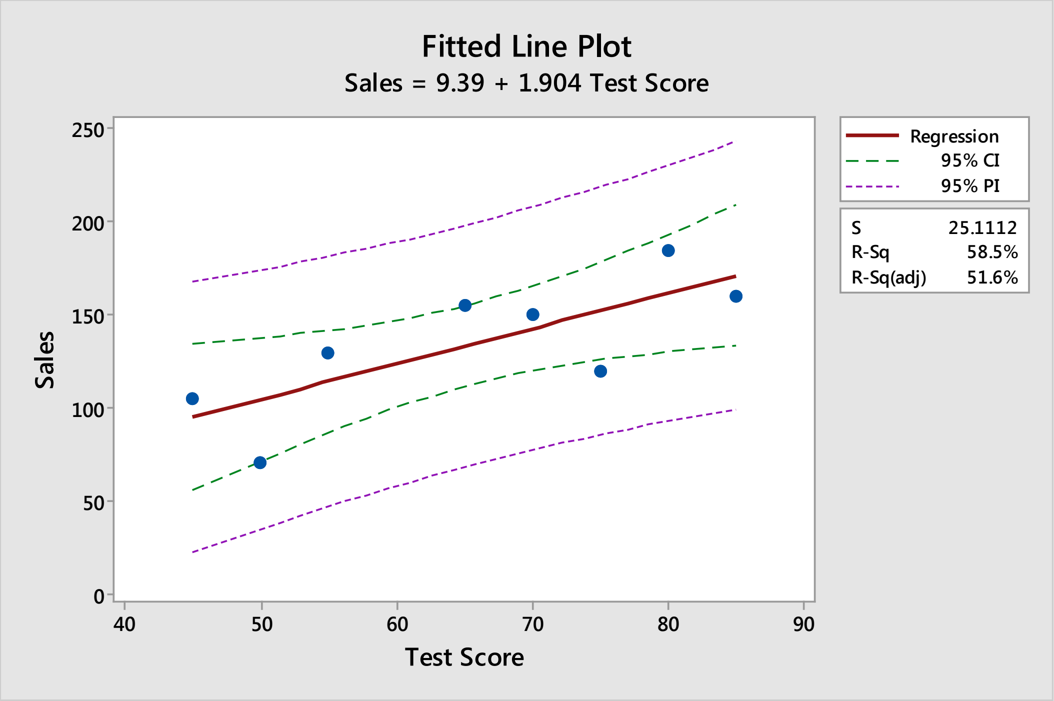
Step 6 Validate Assumptions



Step 6 Predictions using the Model

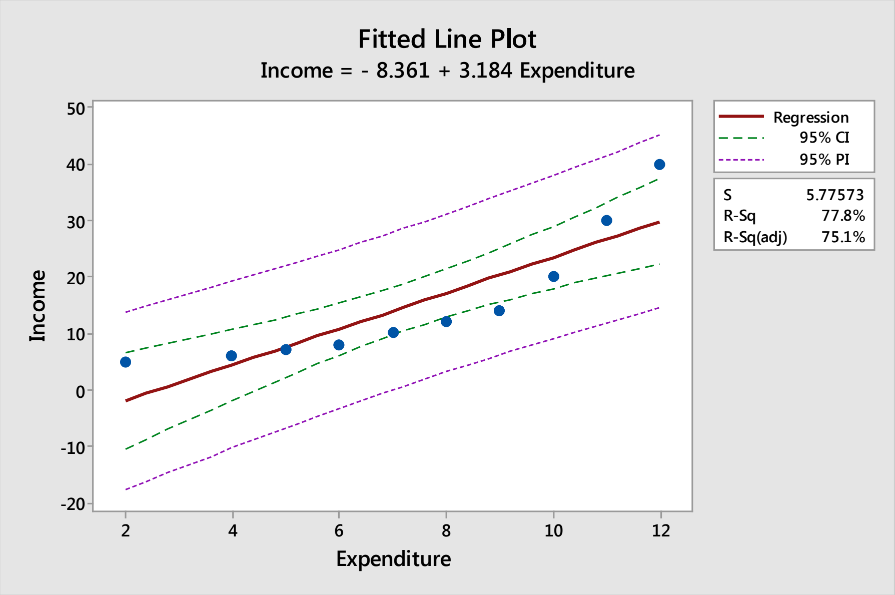
Prediction for Sales

```
Regression Equation

Sales = 9.4 + 1.904 Test Score


Variable    Setting
Test Score       60


    Fit    SE Fit        95% CI                 95% PI
123.662  9.61129  (100.144, 147.180)  (57.8705, 189.454)
```


Fitted Line Plot
Sales = 9.39 + 1.904 Test Score
Sales
Test Score
250
200
150
100
50
0
40
50
60
70
80
90
Regression
95% CI
95% PI
S 25.1112
R-Sq 58.5%
R-Sq(adj) 51.6%

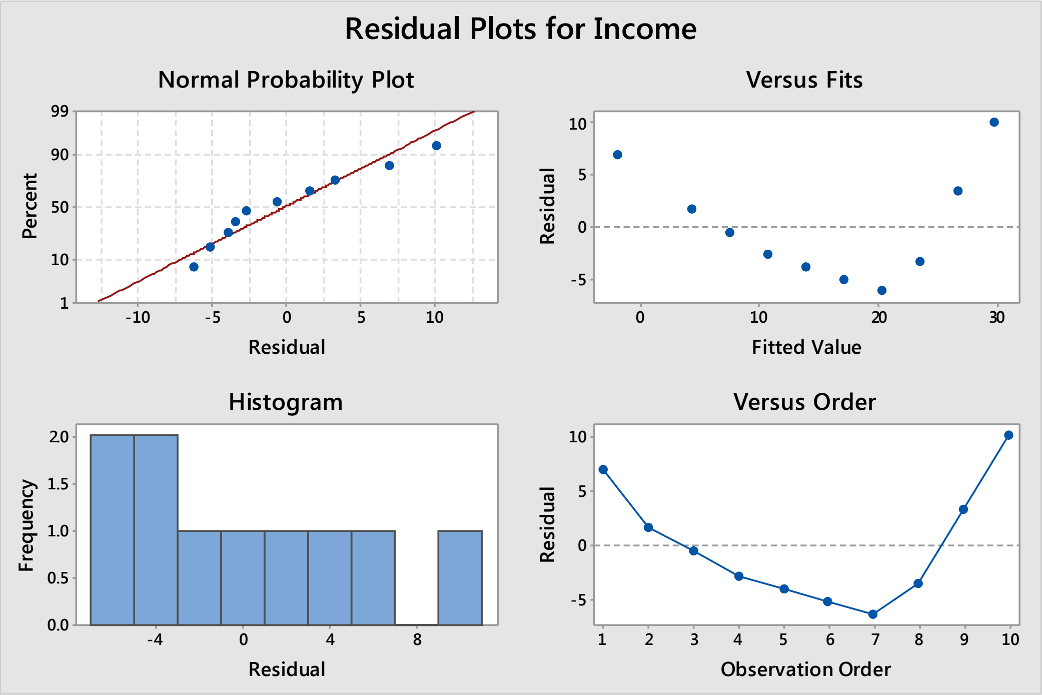
Example 2 : Income vs Expenditure
: Can we predict Income from Expenditure?

Expenditure	Income
2	5
4	6
5	7
6	8
7	10
8	12
9	14
10	20
11	30
12	40



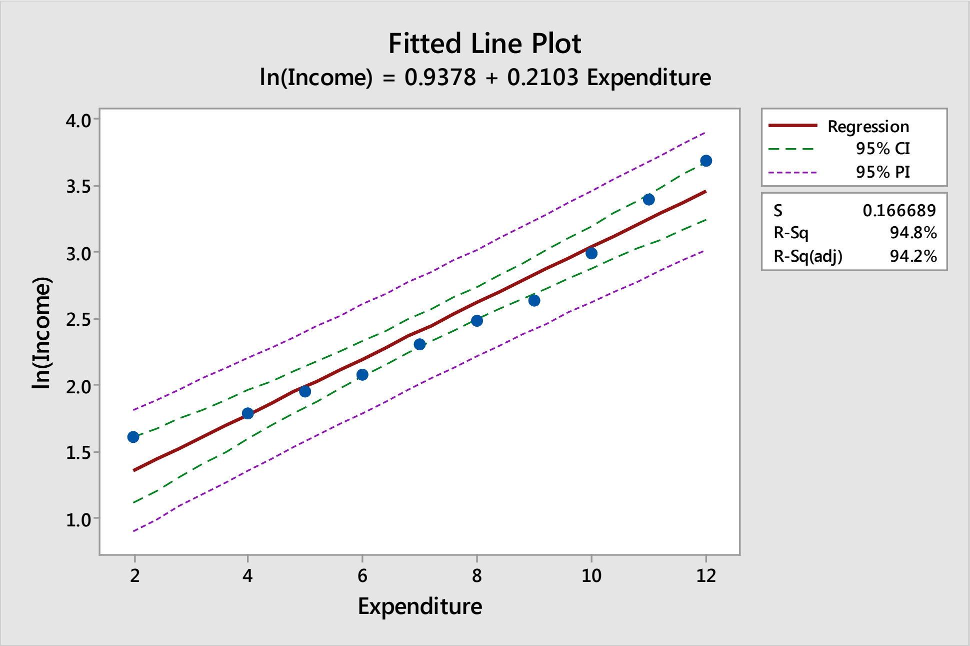
Regression Analysis: Income versus Expenditure

```
Analysis of Variance
Source         DF  Adj SS  Adj MS  F-Value  P-Value
Regression      1   936.7  936.73    28.08    0.001
  Expenditure   1   936.7  936.73    28.08    0.001
Error           8   266.9   33.36
Total           9  1203.6

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
5.77573  77.83%     75.06%      56.18%


Coefficients
Term          Coef  SE Coef  T-Value  P-Value   VIF
Constant     -8.36     4.81    -1.74    0.120
Expenditure  3.184    0.601     5.30    0.001  1.00

Regression Equation

Income = -8.36 + 3.184 Expenditure

Fits and Diagnostics for Unusual Observations
                             Std
Obs  Income    Fit  Resid  Resid
 10   40.00  29.85  10.15   2.15  R

R  Large residual
```



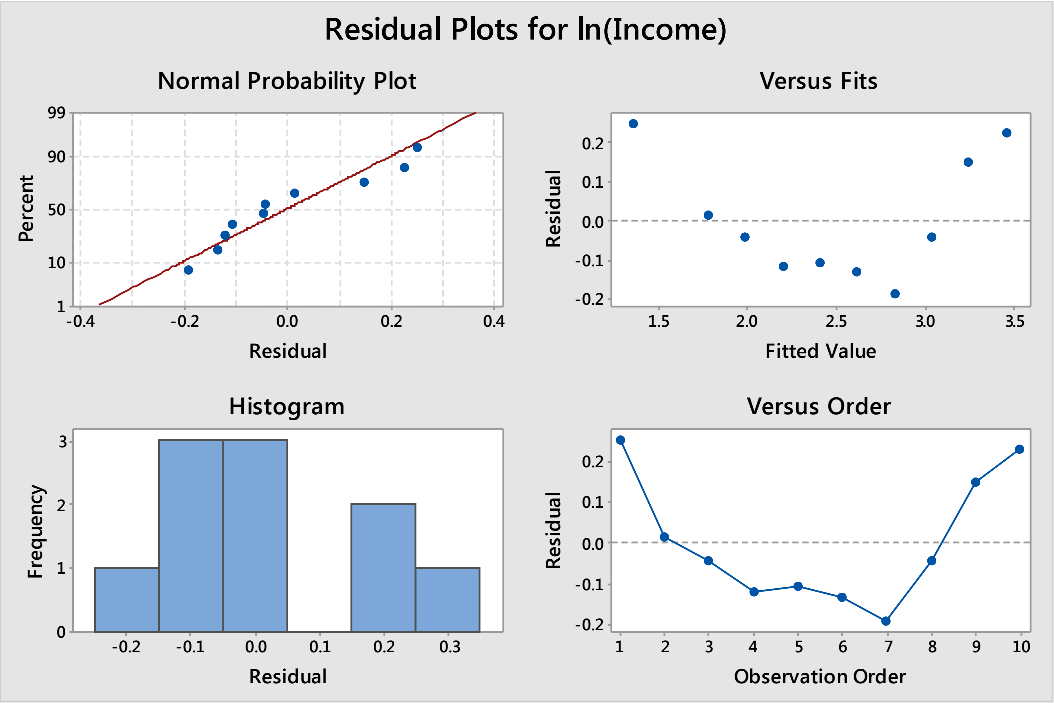
Example 2 (part 2!!)

Expenditure	Income	ln(Income)
2	5	1.60944
4	6	1.79176
5	7	1.94591
6	8	2.07944
7	10	2.30259
8	12	2.48491
9	14	2.63906
10	20	2.99573
11	30	3.40120
12	40	3.68888



Regression Analysis: ln(Income) versus Expenditure

```
Analysis of Variance

Source          DF   Adj SS    Adj MS  F-Value  P-Value
Regression       1   4.0858   4.08581   147.05    0.000
  Expenditure    1   4.0858   4.08581   147.05    0.000
Error            8   0.2223   0.02779
Total            9   4.3081


Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.166689  94.84%     94.20%      89.59%


Coefficients

Term              Coef  SE Coef  T-Value  P-Value   VIF
Constant         0.938    0.139     6.76    0.000
Expenditure     0.2103   0.0173    12.13    0.000  1.00


Regression Equation

ln(Income) = 0.938 + 0.2103 Expenditure
```


Residual Plots for ln(Income)
Normal Probability Plot
Percent
99
90
50
10
1
-0.4
-0.2
0.0
0.2
0.4
Residual
Versus Fits
Residual
0.2
0.1
0.0
-0.1
-0.2
1.5
2.0
2.5
3.0
3.5
Fitted Value
Histogram
Frequency
3
2
1
0
-0.2
-0.1
0.0
0.1
0.2
0.3
Residual
Versus Order
Residual
0.2
0.1
0.0
-0.1
-0.2
1
2
3
4
5
6
7
8
9
10
Observation Order

7.8 Multiple Linear Regression

In previous lectures we examined simple linear regression which is concerned with the relationship between the value of one variable and the value of another variable in the situation in which this relationship is represented by a straight line.

It is often useful to express the mean value of one variable in terms of not one variable but of several others. Some examples will illustrate some slightly different purposes of this approach.

(i) The primary purpose may be to study the effect on variable Y of changes in a particular single variable X_1, but it may be recognised that Y may be affected by several other variables X_2, X_3, X_4,...... The effect on Y of simultaneous changes in X_1, X_2, X_3, X_4,...... must therefore be studied.

Example

In an analysis on data on respiratory function of workers in a particular industry, the effect of duration of exposure to a hazard may be of primary interest. However, respiratory function is affected by age and age is related to duration of exposure. The simultaneous effect on respiratory function of age and exposure must therefore be studied so that the effect on workers of a fixed age may be estimated.

(ii) One may wish to derive insight into some causative mechanism by discovering which of a set of variables X_1, X_2, X_3, X_4,...... has apparently most influence on a dependent variable.

Example 1

By relating stillbirth rate simultaneously to a large number of variables describing the towns - economic, social, meteorological or demographic, for instance - it may be possible to find which factors exert particular influence on the stillbirth rate.

Example 2

The study of variations in the cost per patient in different hospitals. This presumably depends markedly on the ‘patient’ mix - the proportion of different types of patient admitted - as well as other factors. A study of the simultaneous effect of many such variables may explain much of the variation in hospital costs, and, by drawing attention to particular hospitals whose high or low costs are out of line with prediction, may suggest new factors of importance.

(iii) To predict the value of the dependent variable in future individuals.

Example

After treatment of patients with advanced breast cancer by ablative procedures, prognosis is very uncertain. If future progress can be shown to depend on various variables available at the time of operation, it may be possible to predict which patients have a poor prognosis and to consider alternative methods of treatment for them.

The appropriate technique is called **multiple regression** (we will restrict our attention to the linear case). In general, the approach is to express the mean value of the Y variable, usually called the **dependent** or **response** variable, in terms of the values of a set of other variables, usually called the **independent** or **explanatory** variables.

The general form of the multiple linear regression equation is:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \square\square\square + \beta_k X_k$$

We want to find the values of $\alpha, \beta_1, \beta_2, ...\beta_k$ which give the highest possible correlation coefficient between the observed values and the predicted values of Y. When this is achieved we have found the multiple correlation coefficient, R.

Least Squares Estimates of $\alpha, \beta_1, \beta_2, ...\beta_k$

As with the case of a single X variable, the total sum of squares for the Y variable is

$$\sum\left(Y_i - \bar{Y}\right)^2 = \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y}_i - \bar{Y}\right)^2$$

	Total Variation	=	Unexplained Variation	+	Explained Variation
i.e.	Total SS	=	Residual SS	+	Regression SS

We want to find the values of $\alpha, \beta_1, \beta_2, ...\beta_k$ which minimise the residual sum of squares , $\sum\left(Y_i - \hat{Y}_i\right)^2$.

The best estimate of α, denoted by $\hat{\alpha}$**,** is given by

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 - \square\square - \hat{\beta}_k \bar{X}_k$$

<u>Example with two explanatory variables</u>

A patient’s blood pressure modelled in terms of his age and weight

The regression equation is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

As before

$$S_{YY} = \sum \left(Y - \bar{Y}\right)^2$$

$$S_{X_1 X_1} = \sum \left(X_1 - \bar{X}_1\right)^2$$

$$S_{X_2 X_2} = \sum \left(X_2 - \bar{X}_2\right)^2$$

$$S_{X_1 Y} = \sum \left(X_1 - \bar{X}_1\right)\left(Y - \bar{Y}\right)$$

$$S_{X_2 Y} = \sum \left(X_2 - \bar{X}_2\right)\left(Y - \bar{Y}\right)$$

$$S_{X_1 X_2} = \sum \left(X_1 - \bar{X}_1\right)\left(X_2 - \bar{X}_2\right)$$

The best estimates of β_1 *and* β_2 are given by:

$$\hat{\beta}_1 = \frac{S_{X_1 Y} S_{X_2 X_2} - S_{X_2 Y} S_{X_1 X_2}}{S_{X_1 X_1} S_{X_2 X_2} - \left[S_{X_1 X_2}\right]^2}$$

and $$\hat{\beta}_2 = \frac{S_{X_2 Y} S_{X_1 X_1} - S_{X_1 Y} S_{X_1 X_2}}{S_{X_1 X_1} S_{X_2 X_2} - \left[S_{X_1 X_2}\right]^2}$$

The best estimate of α is given by : $\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$

The meaning of a regression coefficient in multiple regression is as follows :

If all variables except X_i are held constant then β_i is the amount by which Y increases with a unit increase in X_i

The regression sum of squares, SS_{reg} is given by

$$SS_{reg} \quad = \quad \sum\left(\hat{Y} - \bar{Y}\right)^2 \quad = \quad \hat{\beta}_1 S_{X_1Y} \quad + \quad \hat{\beta}_2 S_{X_2Y} \quad + \square\square + \quad \hat{\beta}_k S_{X_kY}$$

Since $\qquad SS_{tot} \quad = \quad SS_{reg} \quad + \quad SS_{res}$

then $\qquad SS_{res} \quad = \quad SS_{tot} \quad - \quad SS_{reg}$

$$= \quad \sum\left(Y - \bar{Y}\right)^2 \quad - \quad \sum\left(\hat{Y} - \bar{Y}\right)^2$$

$$= \quad S_{YY} \quad - \quad \left[\hat{\beta}_1 S_{X_1Y} \quad + \quad \hat{\beta}_2 S_{X_2Y} \quad + \square\square + \quad \hat{\beta}_k S_{X_kY}\right]$$

The square of the multiple correlation coefficient which is known as the **Coefficient of Determination** and denoted by R^2 is given by

$$R^2 \quad = \quad \frac{Explained\ \ Variation}{Total\ Variation}$$

$$= \quad \frac{SS_{reg}}{SS_{tot}}$$

The **Multiple Correlation Coefficient, R**, is

$$R \quad = \quad \sqrt{\frac{SS_{reg}}{SS_{tot}}}$$

The Standard Error of β_i

The standard error of a regression coefficient β_i when several X variables are involved is given by

$$s_{\beta_{ii}} \quad = \quad \sqrt{\frac{MS_{res}}{S_{X_i X_{ii}} \left(1 - R_i^2\right)}}$$

where R_i^2 is the squared multiple correlation coefficient of X_i and the remaining X variables and

$$MS_{res} \quad = \quad \frac{SS_{res}}{n - k - 1}$$

Note In the case of two X variables $R_i^2 = r^2$ where r is the correlation of X_1 and X_2

Thus the **95% Confidence Interval for $\boldsymbol{\beta_i}$** is given by

$$\hat{\beta}_i \quad \pm \quad t(n - k - 1; 0.025)\, s_{\beta_i}$$

Test of Significance of β_1 and β_2

For a test of significance of a regression coefficient in a multiple linear regression we a have t test which is essentially the same as the one we had for the regression with one X variable.

Hypotheses $\quad H_o \; : \; \beta_i \; = \; 0$

$\qquad\qquad H_1 \; : \; \beta_i \; \neq \; 0$

Significance Level 0.05

Test Statistic $\quad t \; = \; \dfrac{\hat{\beta}_i}{s_{\beta_i}}$

Rejection Region t(n-k-1; 0.025)

The test of β_i is simply a test to determine whether the regression sum of squares already accounted for by including the other X variables is increased significantly by including the X_i variable as the last variable in the model.

Test of Significance of R

A test for the null hypothesis that R = 0 which is equivalent to testing the null hypothesis that β_1 and β_2 are both equal to zero is given by

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

or

$$F = \frac{SS_{reg}/k}{SS_{res}/(n-k-1)} = \frac{MS_{reg}}{MS_{res}}$$

Under Ho , this quantity follows a F distribution with k and n-k-1 degrees of freedom i.e. The rejection region for a two-sided test at 5% significance

Checking the Assumptions

The mathematical model is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \square\square + \beta_k X_k + \varepsilon$$

$$\textit{where} \quad \varepsilon \sim N(0,\sigma)$$

i.e. for given X variables the values of Y vary about the regression surface in a Normal distribution with mean 0 and standard deviation σ.

Checks

1. Plot the Residuals versus Fitted Values
   This checks that σ^2 is constant for varying Y.

2. Plot Residuals versus Explanatory Variables.
   This checks linearity assumption.

3. Check for the Normality of Residuals : Normal Probability plot of the Residuals

Summary

So far we have concentrated on one fairly simple case, with only two explanatory variables. In this case we investigated several aspects of the data as follows :

(I) Estimated the coefficients and provided corresponding interval estimates

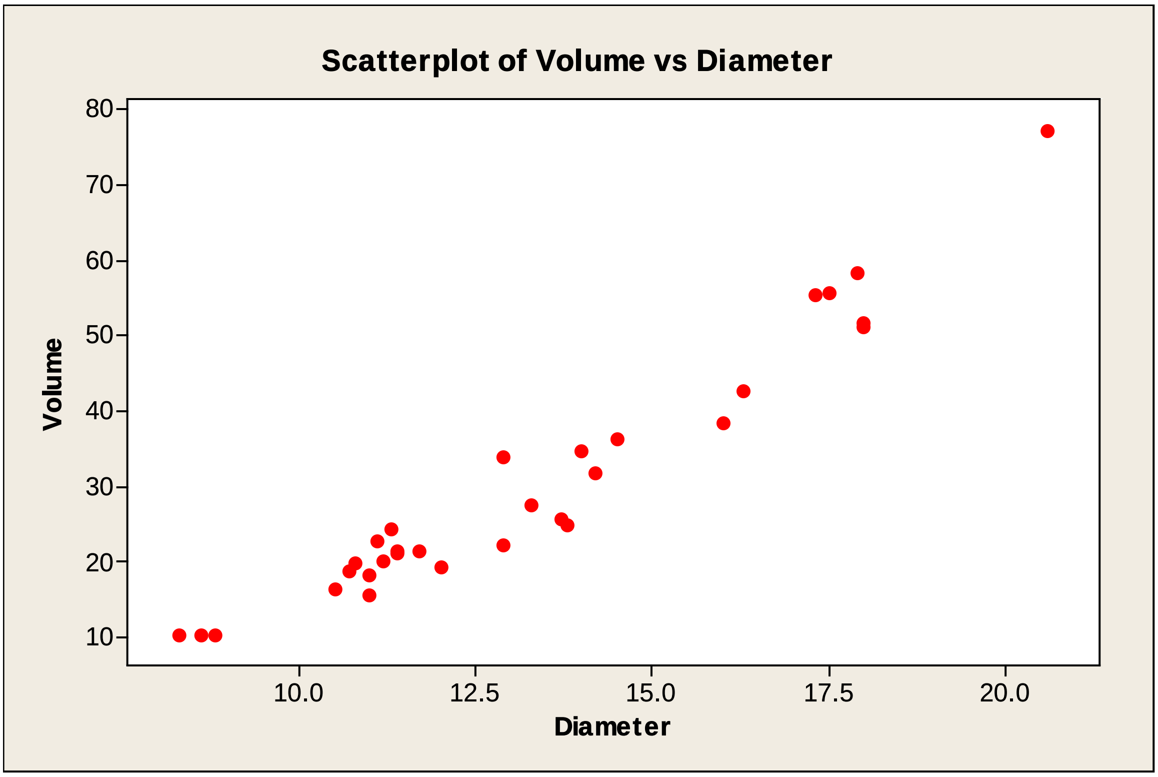
(II) Calculated the coefficient of determination, R^2, and the multiple correlation coefficient, R.

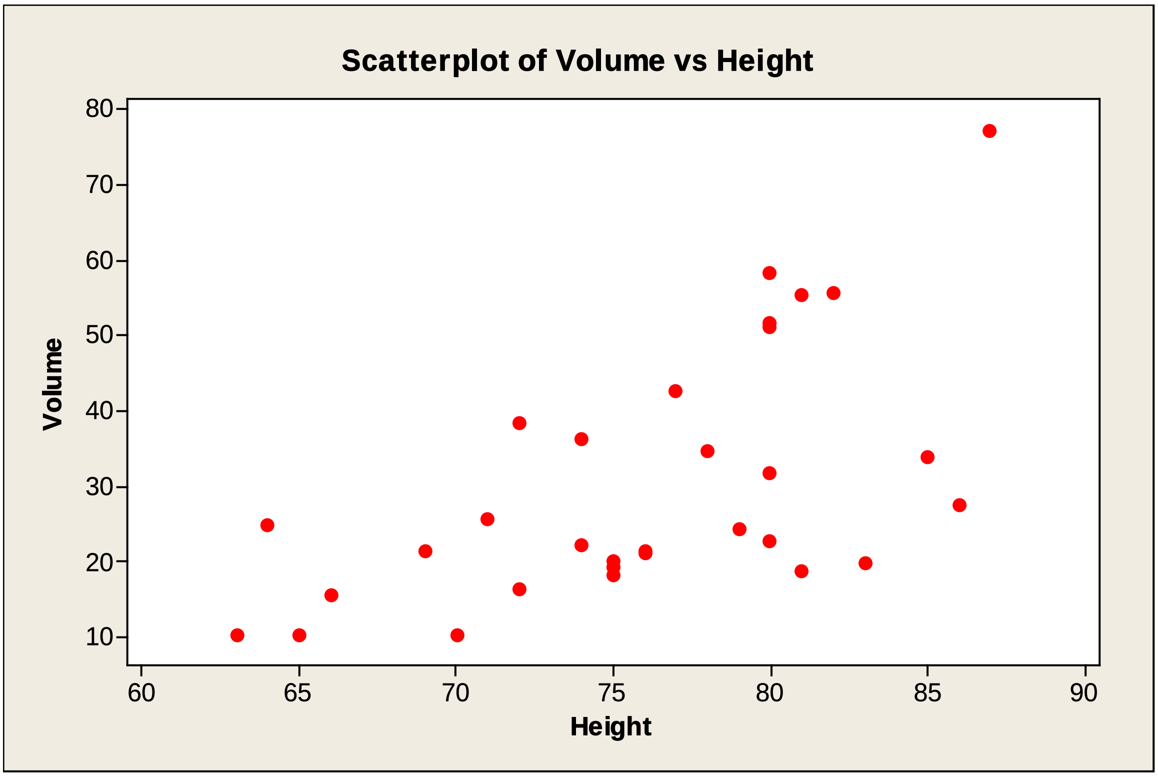
(III) Tested whether a particular explanatory variable significantly improves the model.

(IV) Carried out a test which considers two equivalent properties

- (a) Are β_1 and β_2 both zero?
- (b) Is R zero?

Example Can we predict Volume from Diameter &/or Height?





Correlations: Diameter, Height, Volume

```
          Diameter    Height
Height       0.519
             0.003

Volume       0.967     0.598
             0.000     0.000


Cell Contents: Pearson correlation
               P-Value
```

Regression Analysis: Volume versus Diameter

```
The regression equation is
Volume = - 36.9 + 5.07 Diameter


Predictor     Coef  SE Coef       T      P
Constant   -36.943    3.365  -10.98  0.000
Diameter    5.0659   0.2474   20.48  0.000


S = 4.25199   R-Sq = 93.5%   R-Sq(adj) = 93.3%


Analysis of Variance

Source          DF      SS      MS        F      P
Regression       1  7581.8  7581.8  419.36  0.000
Residual Error  29   524.3    18.1
Total           30  8106.1


Unusual Observations

Obs  Diameter  Volume     Fit  SE Fit  Residual  St Resid
 31      20.6  77.000  67.413   1.972     9.587      2.55RX

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage
```

Regression Analysis: Volume versus Height

```
The regression equation is
Volume = - 87.1 + 1.54 Height


Predictor    Coef  SE Coef      T      P
Constant   -87.12    29.27  -2.98  0.006
Height     1.5433   0.3839   4.02  0.000


S = 13.3970   R-Sq = 35.8%   R-Sq(adj) = 33.6%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  2901.2  2901.2  16.16  0.000
Residual Error  29  5204.9   179.5
Total           30  8106.1


Unusual Observations

Obs  Height  Volume    Fit  SE Fit  Residual  St Resid
 31    87.0   77.00  47.15    4.86     29.85      2.39R

R denotes an observation with a large standardized residual.
```

Regression Analysis: Volume versus Diameter, Height

```
The regression equation is
Volume = - 58.0 + 4.71 Diameter + 0.339 Height


Predictor     Coef  SE Coef      T      P
Constant   -57.988    8.638  -6.71  0.000
Diameter    4.7082   0.2643  17.82  0.000
Height      0.3393   0.1302   2.61  0.014


S = 3.88183   R-Sq = 94.8%   R-Sq(adj) = 94.4%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       2  7684.2  3842.1  254.97  0.000
Residual Error  28   421.9    15.1
Total           30  8106.1


Source    DF  Seq SS
Diameter   1  7581.8
Height     1   102.4


Unusual Observations

Obs  Diameter  Volume     Fit  SE Fit  Residual  St Resid
 31      20.6  77.000  68.515   1.850     8.485      2.49R

R denotes an observation with a large standardized residual.
```

Multiple Linear Regression (continued)

In this section we are going to discuss the following topics related to finding the 'Best Model to Predict a Dependent Variable Y'.

Best Model in this context means 'Highest R^2 but with as few independent variables as possible'.

1. Stepwise Regression
   - 1.1 Forward Stepping
   - 1.2 Backward Stepping

2. Stopping Rules
   - 2.1 Using R^2
   - 2.2 Using Adjusted R^2

3. Multicollinearity

4. Validating the Model
   - 4.1 Splitting the Data into a Training Set and a Test Set
   - 4.2 Using the Training Set to develop a model
   - 4.3 Using the Test Set to 'test' the model

5. Drawbacks of Stepwise Regression

6. Example using Minitab

Multiple Linear Regression (continued)

In this section we are going to discuss the following topics related to finding the 'Best Model to Predict a Dependent Variable Y'.

Best Model in this context means

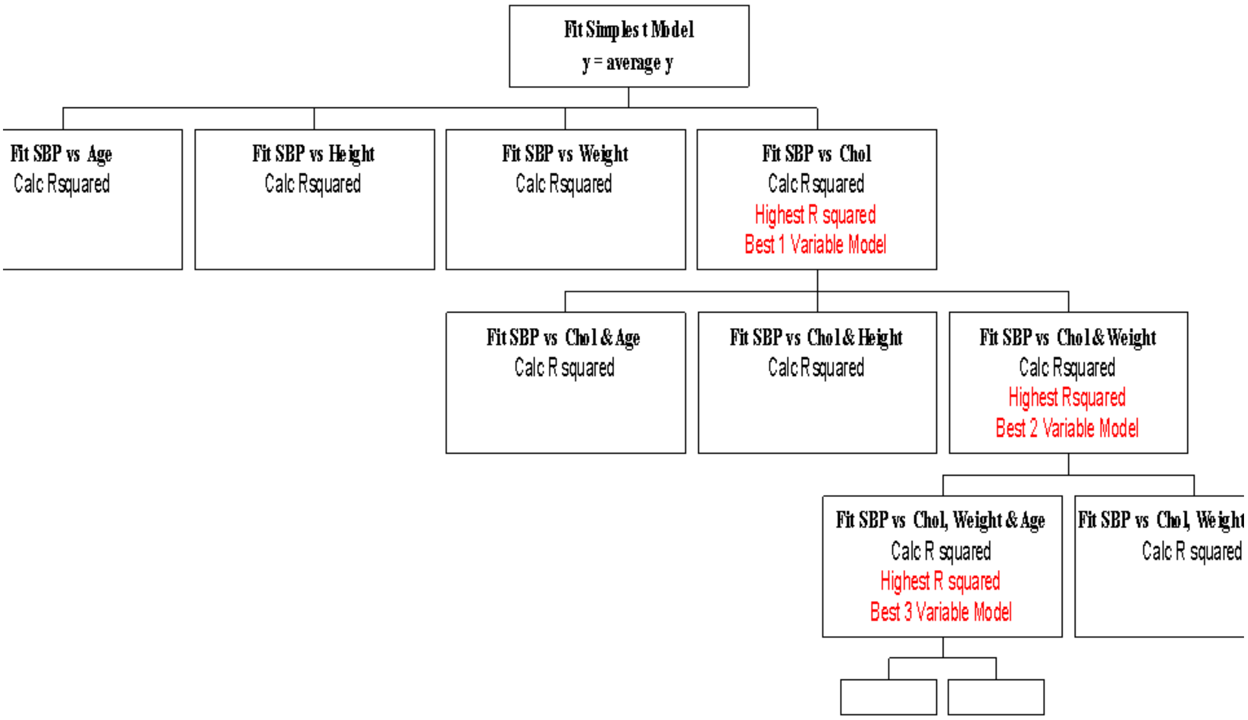
'Highest R^2 but with as few independent variables as possible'.

1. Stepwise Regression

1.1 Forward Stepping
Start with Simplest Model and add one variable in at a time

1.2 Backward Stepping
Start with Most Complex Model and omit one variable at a time.

Find the 'Best Model' for SBP using Age, Height, Weight, Cholesterol....
Using Forward Stepping

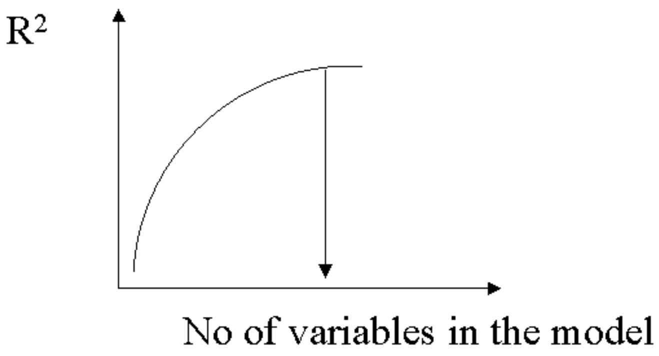


2. Stopping Rules

2.1 Using R^2

At each stage, the Best Model has the highest R^2.As you add another variable into the model R^2 will necessarily increase even if only slightly.

So choose the 'Best Model' when R^2 'levels off'



Important

In addition, check that for the ‘Best Model’ at each stage all variables are ‘important’

i.e. check that the slopes are all significantly different to zero.

2.2 Using Adjusted R^2

Alternatively, find the Maximum Adjusted R^2

This adjusts for the number of variables in the model so does not necessarily increase as you add in new variables. In fact it will reach a maximum and then fall.

$$R_a^2 \quad = \quad 1 - \quad \frac{n-1}{n-m-1}\left(1-R^2\right)$$

n = number of data points, m = number of variables in the model and R^2 is as usual

3. Multicollinearity

Multicollinearity is the correlation between the independent X variables.

No problem if it is slight.

If the any of the pairwise correlations are high then the conclusions related the the highly correlated X variables may be spurious.
Can result in unreasonable regression coefficients.

Check the correlations between the X variables before starting!!

4. Validating the Model

4.1 Splitting the Data into a Training Set and a Test Set

‘Randomly’ split the data into two groups.
For example, toss a coin or use random number generator.
If data set is small, use most data to ‘Train’ the model

4.2 Using the Training Set to develop a model

Use Forward Stepping on this part of the data
Check the assumptions for the ‘Best Model’ and note the regression equation.

4.3 Using the Test Set to 'test' the model

(i) Using the regression equation from the Training Set - predict the dependent variable

i.e. use the coefficients from the Training Set but the data from the Test Set

i.e. We have the ‘Predicted’ values

(ii) Compare the ‘predicted’ with the ‘observed’

(a) Calculate the mean and sd

(b) Do a paired t test

5. Drawbacks of Stepwise Regression

We may miss the ‘best model’ because we do not investigate all combinations of variables.

We must be prepared to ‘live with’ that loss!!!

7.9 Example Using Minitab

We want to find the ‘Best Model’ to predict Systolic Blood Pressure using the following variables :

Age, Years, Weight, Height, Chin, Forearm, Calf, Pulse and Diastolic Blood Pressure

Fitting all independent variables

Regression Analysis: Systol versus Age, Years, Weight, Height, Chin, ...

```
Analysis of Variance

Source      DF   Adj SS   Adj MS  F-Value  P-Value
Regression   9  2982.95   331.44     2.28    0.081
  Age        1   180.08   180.08     1.24    0.285
  Years      1   257.52   257.52     1.77    0.205
  Weight     1  1249.92  1249.92     8.59    0.011
  Height     1   215.27   215.27     1.48    0.244
  Chin       1   116.86   116.86     0.80    0.385
  Forearm    1    42.58    42.58     0.29    0.597
  Calf       1    26.96    26.96     0.19    0.673
  Pulse      1     9.97     9.97     0.07    0.797
  Diastol    1   235.04   235.04     1.62    0.224
Error       14  2037.01   145.50
Total       23  5019.96


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
12.0624  59.42%     33.34%       0.00%


Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant    115.4     86.9     1.33    0.206
Age        -0.522    0.469    -1.11    0.285  2.30
Years      -0.463    0.348    -1.33    0.205  2.11
Weight      2.224    0.759     2.93    0.011  2.87
Height    -0.0780   0.0641    -1.22    0.244  1.85
Chin        -1.46     1.63    -0.90    0.385  2.34
Forearm     -1.09     2.02    -0.54    0.597  3.31
Calf         0.44     1.01     0.43    0.673  2.87
Pulse       0.089    0.338     0.26    0.797  1.42
Diastol     0.305    0.240     1.27    0.224  1.35


Regression Equation

Systol = 115.4 - 0.522 Age - 0.463 Years + 2.224 Weight - 0.0780 Height
         - 1.46 Chin - 1.09 Forearm + 0.44 Calf + 0.089 Pulse + 0.305 Diastol


Fits and Diagnostics for Unusual Observations

                              Std
Obs  Systol     Fit  Resid  Resid
 16  170.00  151.99  18.01   2.34  R

R  Large residual
```

Training Set

Age	Years	Weight	Height	Chin	Forearm	Calf	Pulse	Systol	Diastol	Random Nos.
21	1	71	1629	8	7	12.7	88	170	76	1
22	6	56.5	1569	3.3	5	8	64	120	60	1
28	5	53	1494	7.3	4.7	8	64	120	76	1
28	25	53	1568	3.7	4.3	0	80	108	62	1
32	13	57	1530	5.7	4	6	60	134	64	1
33	13	66.5	1622	6	5.7	8.3	68	116	76	1
33	10	59.1	1486	6.7	5.3	10.3	72	114	74	1
34	15	64	1578	3.3	5.3	7	88	130	80	1
35	18	69.5	1645	9.3	5	7	60	118	68	1
35	2	64	1648	3	3.7	6.7	60	138	78	1
36	12	56.5	1521	3.3	5	11.7	72	134	86	1
36	15	57	1547	3	3	6	84	120	70	1
37	16	55	1505	4.3	5	7	64	120	76	1
38	10	58	1538	8.7	6	13	64	124	64	1
38	18	59.5	1513	5.3	4	7.7	80	114	66	1
38	11	61	1653	4	3.3	4	76	136	78	1
39	21	57.5	1580	4	3	5	64	124	62	1
39	24	74	1647	7.3	6.3	15.7	64	128	84	1
39	14	72	1620	6.3	7.7	13.3	68	134	92	1
42	12	68	1605	11	7	10.7	88	128	90	1
43	26	73	1615	12	4	5.7	68	138	74	1
45	10	60.2	1534	3	3	3.3	56	134	70	1
47	1	55	1536	3	3	4	64	116	54	1
54	40	87	1542	11.3	11.7	11.3	92	152	88	1

Test Set

Age	Years	Weight	Height	Chin	Forearm	Calf	Pulse	Systol	Diastol	Random Nos.
24	5	56	1561	3.3	1.3	4.3	68	125	75	0
24	1	61	1619	3.7	3	4.3	52	148	120	0
25	1	65	1566	9	12.7	20.7	72	140	78	0
27	19	62	1639	3	3.3	5.7	72	106	72	0
31	6	65	1540	10.3	9	10	76	124	70	0
37	17	57	1473	6	5.3	11.7	72	114	80	0
38	11	57	1566	3	3	3	60	126	72	0
41	25	62.5	1637	6	5.3	8	76	112	80	0
41	32	68	1528	10	5	11.3	60	128	82	0
41	5	63.4	1647	5.3	4.3	13.7	76	134	92	0
43	25	69	1625	5	3	6	72	140	72	0
43	10	64	1640	5.7	3	7	60	118	66	0
44	19	65	1610	8	6.7	7.7	74	110	70	0
44	18	71	1572	3	4.7	4.3	72	142	84	0
50	43	70	1630	4	6	11.7	72	132	90	0

Step 1:

Correlation: Systol, Age, Years, Weight, Height, Chin, Forearm, Calf, ...

```
          Systol      Age    Years   Weight   Height     Chin  Forearm     Calf
Age        0.069
Years      0.008    0.482
Weight     0.596    0.413    0.518
Height     0.344   -0.057    0.033    0.520
Chin       0.313    0.258    0.415    0.676    0.191
Forearm    0.443    0.204    0.426    0.744    0.089    0.619
Calf       0.366    0.002    0.017    0.501    0.109    0.463    0.678
Pulse      0.305    0.037    0.304    0.344    0.022    0.229    0.472    0.152
Diastol    0.456    0.228    0.231    0.607    0.280    0.365    0.620    0.589

           Pulse
Diastol    0.377

Cell Contents: Pearson correlation
```

Step 2 : Find the Best Model

METHOD 1:

Regression Analysis: Systol versus Age, Years, Weight, Height, Chin, ...

```
Stepwise Selection of Terms

Candidate terms: Age, Years, Weight, Height, Chin, Forearm, Calf, Pulse,
     Diastol

             ----Step 1----    -----Step 2----
              Coef        P      Coef        P
Constant      67.2               54.3
Weight       0.967    0.002     1.313    0.000
Years                          -0.626    0.037

S                   11.1251           10.2386
R-sq                 35.47%            47.83%
R-sq(adj)            32.54%            42.86%
R-sq(pred)           24.83%            20.89%
Mallows' Cp           -0.30             -2.07

α to enter = 0.15, α to remove = 0.15


Analysis of Variance

Source      DF  Adj SS  Adj MS  F-Value  P-Value
Regression   2  2018.4  1009.2     9.63    0.001
  Years      1   521.5   521.5     4.97    0.037
  Weight     1  2018.2  2018.2    19.25    0.000
Error       21  2201.4   104.8
Total       23  4219.8


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
10.2386  47.83%     42.86%      20.89%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant    54.3     17.2     3.16    0.005
Years     -0.626    0.280    -2.23    0.037  1.37
Weight     1.313    0.299     4.39    0.000  1.37


Regression Equation

Systol = 54.3 - 0.626 Years + 1.313 Weight
```

METHOD 1a:

Regression Analysis: Systol versus Age, Years, Weight, Height, Chin, ...

```
Backward Elimination of Terms

Candidate terms: Age, Years, Weight, Height, Chin, Forearm, Calf, Pulse,
     Diastol

             -----Step 1-----    -----Step 2-----    -----Step 3-----
                Coef        P       Coef        P       Coef        P
Constant         119                114.0               112.0
Age           -0.339    0.482     -0.308    0.501     -0.283    0.520
Years         -0.685    0.099     -0.636    0.092     -0.658    0.070
Weight         1.920    0.042      1.888    0.037      1.915    0.029
Height       -0.0643    0.422    -0.0604    0.432    -0.0566    0.443
Chin           -0.57    0.656      -0.63    0.604      -0.68    0.564
Forearm        -1.21    0.683      -1.61    0.545      -1.30    0.593
Calf           -0.40    0.725
Pulse          0.161    0.579      0.191    0.479      0.207    0.425
Diastol        0.168    0.654      0.119    0.723

S                     11.7551             11.4087             11.0945
R-sq                   54.16%              53.73%              53.33%
R-sq(adj)              24.68%              29.06%              32.91%
R-sq(pred)              0.00%               0.00%               0.00%
Mallows’ Cp             10.00                8.13                6.25

             -----Step 4-----    -----Step 5-----    -----Step 6----
                Coef        P       Coef        P      Coef        P
Constant        89.4                 71.7               39.4
Age           -0.197    0.621
Years         -0.648    0.067     -0.697    0.037    -0.656    0.036
Weight         1.607    0.010      1.495    0.008     1.355    0.002
Height       -0.0344    0.562    -0.0245    0.652
Chin           -0.74    0.518      -0.65    0.557     -0.53    0.614
Forearm
Calf
Pulse          0.174    0.479      0.205    0.379     0.222    0.325
Diastol

S                     10.8626             10.6350            10.4115
R-sq                   52.46%              51.76%             51.19%
R-sq(adj)              35.69%              38.35%             40.92%
R-sq(pred)              0.00%               3.45%             10.85%
Mallows’ Cp              4.52                2.73               0.90

             -----Step 7----    -----Step 8----
               Coef        P      Coef        P
Constant       43.6               54.3
Age
Years        -0.672    0.028    -0.626    0.037
Weight        1.240    0.001     1.313    0.000
Height
Chin
Forearm
Calf
Pulse         0.224    0.310
Diastol

S                    10.2179            10.2386
R-sq                  50.52%             47.83%
R-sq(adj)             43.09%             42.86%
R-sq(pred)            15.29%             20.89%
Mallows’ Cp            -0.89              -2.07

α to remove = 0.1
```

```
Analysis of Variance

Source      DF  Adj SS  Adj MS  F-Value  P-Value
Regression   2  2018.4  1009.2     9.63    0.001
  Years      1   521.5   521.5     4.97    0.037
  Weight     1  2018.2  2018.2    19.25    0.000
Error       21  2201.4   104.8
Total       23  4219.8


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
10.2386  47.83%     42.86%      20.89%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant    54.3     17.2     3.16    0.005
Years     -0.626    0.280    -2.23    0.037  1.37
Weight     1.313    0.299     4.39    0.000  1.37


Regression Equation

Systol = 54.3 - 0.626 Years + 1.313 Weight
```

METHOD 1b:

Regression Analysis: Systol versus Age, Years, Weight, Height, Chin, ...

```
Forward Selection of Terms

Candidate terms: Age, Years, Weight, Height, Chin, Forearm, Calf, Pulse,
     Diastol

             ----Step 1----    -----Step 2----
              Coef        P      Coef        P
Constant      67.2                54.3
Weight       0.967    0.002     1.313    0.000
Years                          -0.626    0.037

S                  11.1251           10.2386
R-sq                35.47%            47.83%
R-sq(adj)           32.54%            42.86%
R-sq(pred)          24.83%            20.89%
Mallows' Cp          -0.30             -2.07

α to enter = 0.25


Analysis of Variance

Source      DF  Adj SS  Adj MS  F-Value  P-Value
Regression   2  2018.4  1009.2     9.63    0.001
  Years      1   521.5   521.5     4.97    0.037
  Weight     1  2018.2  2018.2    19.25    0.000
Error       21  2201.4   104.8
Total       23  4219.8


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
10.2386  47.83%     42.86%      20.89%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant    54.3     17.2     3.16    0.005
Years     -0.626    0.280    -2.23    0.037  1.37
Weight     1.313    0.299     4.39    0.000  1.37


Regression Equation

Systol = 54.3 - 0.626 Years + 1.313 Weight
```

METHOD 2:

Best Subsets Regression: Systol versus Age, Years, ...

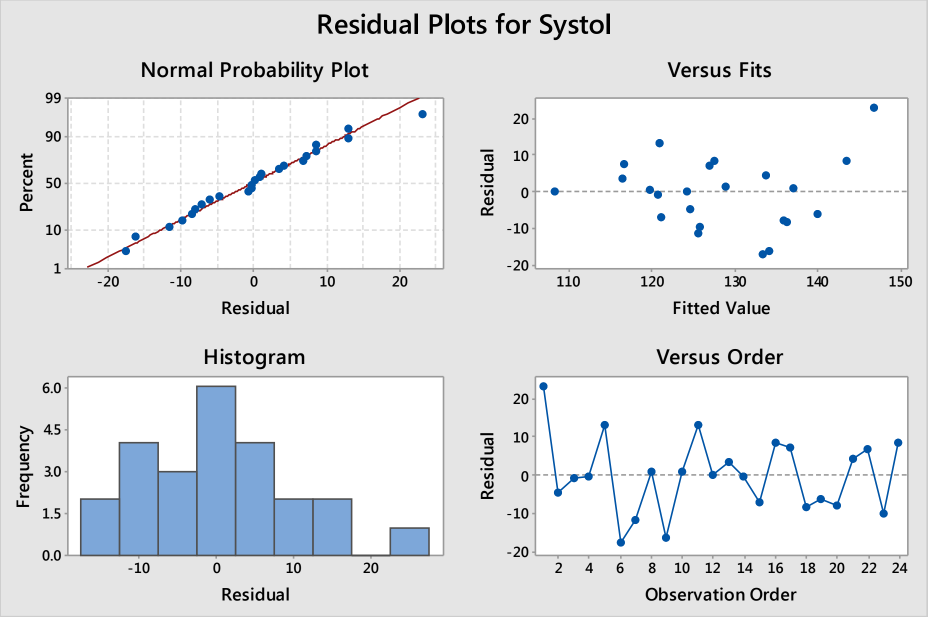
```
Response is Systol

                                                      F     D
                                                  W H o     i
                                                Y e e r   P a
                                                e i i C e C u s
                                              A a g g h a a l t
             R-Sq    R-Sq  Mallows            g r h h i r l s o
Vars  R-Sq  (adj)  (pred)       Cp       S  e s t t n m f e l
   1  35.5   32.5    24.8     -0.3  11.125      X
   1  20.8   17.2    10.0      4.2  12.323                  X
   2  47.8   42.9    20.9     -2.1  10.239    X X
   2  39.2   33.4     0.0      0.6  11.050  X   X
   3  50.5   43.1    15.3     -0.9  10.218    X X         X
   3  48.6   40.9    18.8     -0.3  10.417    X X   X
   4  51.2   40.9    10.9      0.9  10.411    X X   X     X
   4  50.8   40.4     7.0      1.0  10.454    X X X       X
   5  51.8   38.4     3.5      2.7  10.635    X X X X     X
   5  51.5   38.0     0.0      2.8  10.665  X X X   X     X
   6  52.5   35.7     0.0      4.5  10.863  X X X X X     X
   6  52.3   35.5     0.0      4.6  10.879  X X X X   X   X
   7  53.3   32.9     0.0      6.3  11.094  X X X X X X   X
   7  53.1   32.5     0.0      6.3  11.127  X X X X X   X X
   8  53.7   29.1     0.0      8.1  11.409  X X X X X X   X X
   8  53.6   28.8     0.0      8.2  11.427  X X X X X   X X X
   9  54.2   24.7     0.0     10.0  11.755  X X X X X X X X X
```

BEST MODEL IS :

Check assumptions etc for best model

Regression Analysis: Systol versus Weight, Years

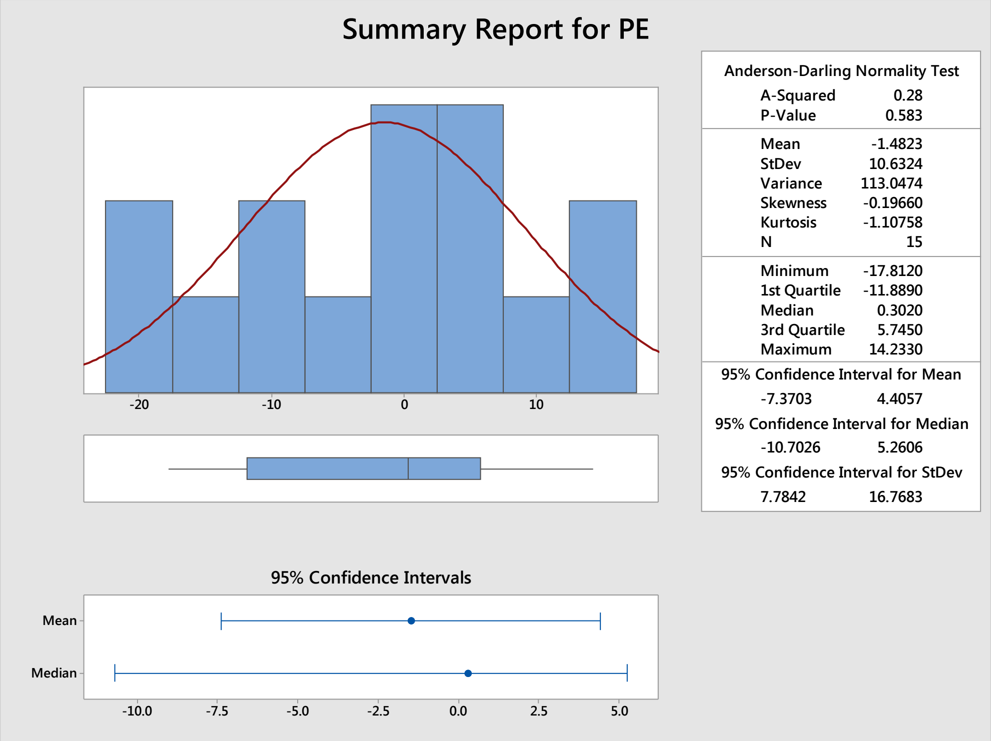
```
Analysis of Variance

Source       DF   Adj SS   Adj MS  F-Value  P-Value
Regression    2   2018.4   1009.2     9.63    0.001
  Years       1    521.5    521.5     4.97    0.037
  Weight      1   2018.2   2018.2    19.25    0.000
Error        21   2201.4    104.8
Total        23   4219.8


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
10.2386  47.83%     42.86%      20.89%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant    54.3     17.2     3.16    0.005
Years     -0.626    0.280    -2.23    0.037  1.37
Weight     1.313    0.299     4.39    0.000  1.37


Regression Equation

Systol = 54.3 - 0.626 Years + 1.313 Weight


Fits and Diagnostics for Unusual Observations

                              Std
Obs  Systol     Fit  Resid  Resid
  1  170.00  146.86  23.14   2.73  R
 24  152.00  143.46   8.54   1.21     X

R  Large residual
X  Unusual X
```



Use model to predict SBP for the Test Set

i.e. Predicted = 54.3 + 1.313 Weight - 0.626 Years

And PE = Observed - Predicted



Summary Report for PE

Anderson-Darling Normality Test

A-Squared	0.28
P-Value	0.583
Mean	-1.4823
StDev	10.6324
Variance	113.0474
Skewness	-0.19660
Kurtosis	-1.10758
N	15
Minimum	-17.8120
1st Quartile	-11.8890
Median	0.3020
3rd Quartile	5.7450
Maximum	14.2330

95% Confidence Interval for Mean

-7.3703 4.4057

95% Confidence Interval for Median

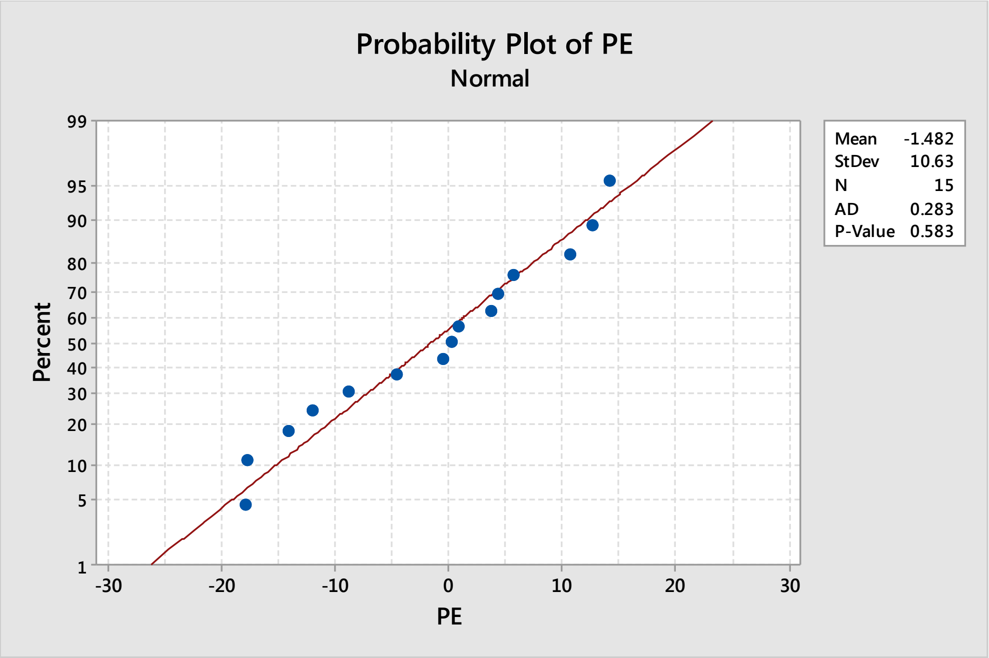
-10.7026 5.2606

95% Confidence Interval for StDev

7.7842 16.7683

Paired T-Test and CI: Systol, Predicted

```
Paired T for Systol - Predicted

             N    Mean  StDev  SE Mean
Systol      15  126.60  12.81     3.31
Predicted   15  128.08   6.79     1.75
Difference  15   -1.48  10.63     2.75


95% CI for mean difference: (-7.37, 4.41)
T-Test of mean difference = 0 (vs ≠ 0): T-Value = -0.54  P-Value = 0.598
```


Probability Plot of PE
Normal
Percent
99
95
90
80
70
60
50
40
30
20
10
5
1
-30
-20
-10
0
10
20
30
PE
Mean -1.482
StDev 10.63
N 15
AD 0.283
P-Value 0.583