BIOSTATISTICS
TOPIC 8: ANALYSIS OF CORRELATIONS
I. SIMPLE LINEAR REGRESSION

Give a man three weapons - correlation, regression and a pen -
and he will use all three.
Anon, 1978

So far we have been concerned with analyses of differences. And, in doing so, we have considered measuring *n* subjects on a single outcome variable (or two groups of *n* subjects on one variable). Such measurements have yielded univariate frequency distribution and the analysis is often referred to as *univariate analysis*. Now, we are considering *n* subjects and in each subject has two measures available; in other words, we have two variables per subject, say *x* and *y*. Our interest in this kind of data is obviously to measure relationship between the two variables. We can plot the value of *y* against the value of *x* in a scatter diagram and assess whether the value of *y* varies systematically with the variation in values of *x*. But we still wants to have a single summary measure of the strength of relationship between *x* and *y*.

In his book "Natural Inheritance", Francis Galton wrote: "each peculiarity in a man is shared by his kinsman, but *on the average*, in a less degree. For example, white tall fathers would tend to have tall sons, the sons would be on the average shorter than their fathers, and sons of short fathers, though having heights below the average for the entire population, would tend to be taller than their fathers." He, then, concluded a phenomenon called "*law of universal regression*" which was the origin of the topic we are learning right now. Today, the characteristics of returning from extreme values toward the average of the full population is well recognised and is termed "regression toward the mean".

We will consider methods for assessing the association between continuous variables using two methods known as **correlation analysis** and **linear regression analysis**, which are happened to be some of the most popular statistical techniques in medical research.

I. CORRELATION ANALYSIS

1.1. THE COVARIANCE AND COEFFICIENT OF CORRELATION

In a previous topic, we stated that if Y and Y are independent variables, then the variance of the sum or difference between X and Y is equal to the variance of X plus the variance of Y, that is:

$$\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y)$$

what happen if X and Y are not independent? Before discussing this problem, we introduce the concepts of *covariance* and *correlation*.

In elementary trigonometry we learn that for a right triangle, if we let the hypotenuse side be c and the other two sides be a and b, the Pythagoras' theorem states that:

$$c^2 = a^2 + b^2$$

and in any triangle:

$$c^2 = a^2 + b^2 - 2ab.\cos C \qquad \text{(Cosine rule).}$$

Analogously, if we have two random variables X and Y, where X may be the height of father and Y may be the height of daughter, their variance can be estimated by:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad [1]$$

and $$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

respectively.

Furthermore, if X and Y are independent, we have:

$$s^2_{X+Y} = s^2_X + s^2_Y \qquad [2]$$

Let us now discuss X and Y in the context of genetics. Let X be BMD of father and Y be the BMD of daughter. It is clear that we can find another expression for the relationship between X and Y by multiplying each father's BMD from its mean $(x_i - \bar{x})$ by corresponding deviation of his daughter $(y_i - \bar{y})$, instead of squaring the father's or daughter's deviation, before summation. We refer this quantity to as covariance between X and Y and is denoted by Cov(X, Y); that is:

$$\mathrm{cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad [3]$$

By definition and analogous to the Cosine law in any triangle, we have: if X and Y are not independent, then:

: $$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2Cov(X,Y) \qquad [4]$$

A number of points need to be noted here:

(a) Variances as defined in [1] are always positive since they are derived from sums of squares, whereas, covariances as defined in [3] are derived from sum of cross-products of deviations and so may be either positive or negative.

(b) A positive value indicates that the deviations from the mean in one distribution, say father's BMDs, are preponderantly accompanied by deviations in the other, say daughter's BMDs, in the same direction, positive or negative.

(c) A negative covariance, on the other hand, indicates that deviations in the two distributions are preponderantly in opposite directions.

(d) When the deviation in one of the distribution is equally likely to be accompanied by deviation of like or opposite sign in the other, the covariance, apart from errors of random sampling, will be zero.

The importance of covariance is now obvious. If variation of BMD is under genetic control we would expect higher BMD fathers generally have high BMD daughters and low

BMD fathers generally have low BMD daughters. In other words, we should expect them to have positive covariance. Lack of genetic control would produce a covariance of zero. It was by this means that Galton first showed stature in man to be under genetic control. He found that the covariance of parent and offspring, and also that of pairs of siblings, was positive.

The size of the covariance relative to some standard gives a measure of the strength of the association between the relatives. The standard taken is that afforded by the variances of the two separate distributions, in our case, of father's BMD and daughter's BMD. We many compare the covariance to these variances separately and we do this by calculating the regression coefficients which have the forms:

$$\frac{Cov(X,Y)}{\text{var}(X)} \quad \text{(regression of daughters on father)}$$

or $$\frac{Cov(X,Y)}{\text{var}(Y)} \quad \text{(regression of fathers on daughters)}$$

we can also compare the covariance with the two variances at once:

$$\frac{Cov(X,Y)}{\sqrt{\text{var}(Y)\times \text{var}(Y)}}$$

This is called the **coefficient of correlation** and is denoted by *r*.

i.e. $$\boxed{r = \frac{Cov(X,Y)}{\sqrt{\text{var}(X).\text{var}(Y)}} = \frac{Cov(X,Y)}{s_x \times s_y}} \qquad [5]$$

r will have a maximum value of |1| (a complete determination of daughter's BMD by father's BMD) and minimum value of 0 (no relationship between father's and daughter's BMDs).

With some algebraic manipulation, we can show that [5] can be written in another way:

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

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} yi\right)}{(n-1)s_x s_y} \qquad [6]$$

where s_x and s_y are standard deviations for X and Y variable, respectively.

1.2. TEST OF HYPOTHESIS

One obvious question is that whether the observed coefficient of correlation (r) is significantly different from zero. Under the null hypothesis that there is no association in the population ($r = 0$), it can be shown that the statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

has a t distribution with n -2 df.

On the other hand, for a moderate or large sample size, we can set up a 95% confidence interval of r by using a theoretical distribution of r. It can be shown that the sampling distribution of r is not normally distributed. We can, however, transform it to a Normal distributed quantity by using the so-called Fisher's transformation in which:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad [7]$$

The standard error of z is approximately equal to:

$$SE(z) = \frac{1}{\sqrt{n-3}} \qquad [8]$$

Thus, approximate 95% confidence interval is:

$$z - \frac{1.96}{\sqrt{n-3}} \quad \text{to} \quad z + \frac{1.96}{\sqrt{n-3}}$$

Of course, we can back-transform the data to obtain 95% confidence interval for r (this is left for exercise).

Example 1: Consider a clinical trial involving patients presenting with hyperlipoproteinaemia, baseline values of the age of patients (years), total serum cholesterol (mg/ml) and serum calcium level (mg/100ml) were recorded. Data for 18 patients are given below:

Patient	Age (X)	Cholesterol (Y)
1	46	3.5
2	20	1.9
3	52	4.0
4	30	2.6
5	57	4.5
6	25	3.0
7	28	2.9
8	36	3.8
9	22	2.1
10	43	3.8
11	57	4.1
12	33	3.0
13	22	2.5
14	63	4.6
15	40	3.2
16	48	4.2
17	28	2.3
18	49	4.0
Mean	38.83	3.33
S.D	13.596	0.838

Let age be X and cholesterol be Y, to calculate the correlation coefficient, we need to calculate the covariance Cov(X, Y) which is:

$$\begin{aligned}\text{Cov}(X,\ Y) &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{xy} \\ &= 10.68\end{aligned}$$

Then the coefficient of correlation is:

$$r = \frac{Cov(X,Y)}{s_x s_y} = \frac{10.68}{13.596 \times 0.838} = 0.937.$$

To test for the significance of r, we need to covert it to the z score as given in [7]:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0.937}{1-0.937}\right) = 0.56$$

with the standard error of z is given in [8]:

$$SE(z) = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{18-3}} = 0.2582$$

Then the t ratio is 0.56 / 0.2582 = 2.165 which exceeds the expected value of 2.11 (with 17 df and 5% significance level), we conclude that there is an association between age and cholesterol in this sample of subjects. //

1.3. TEST FOR DIFFERENCE BETWEEN TWO COEFFICIENTS OF CORRELATION

Suppose that we have two sample coefficients of correlation r_1 and r_2 which were estimated from two unknown population coefficients ρ_1 and ρ_2, respectively. Suppose further that r_1 and r_2 were derived from two independent samples of n_1 and n_2 subjects, respectively. To test the hypothesis that $\rho_1 = \rho_2$ versus the alternative hypothesis that $\rho_1 \neq \rho_2$, we firstly convert these sample coefficients into a z-score:

$$z_1 = \frac{1}{2}\ln\left(\frac{1+r_1}{1-r_1}\right) \qquad \text{and} \qquad z_2 = \frac{1}{2}\ln\left(\frac{1+r_2}{1-r_2}\right)$$

By theory, the statistic $z_1 - z_2$ is distributed about the mean

$$\text{Mean}(z_1 - z_2) = \frac{\rho}{2(n_1 - 1)} - \frac{\rho}{2(n_2 - 1)}$$

where ρ is the common correlation coefficient, with variance

$$\text{Var}(z_1 - z_2) = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$$

If the samples are not small or if n_1 and n_2 are not very different, the statistic

$$t = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

can be used as a test statistic of the hypothesis.

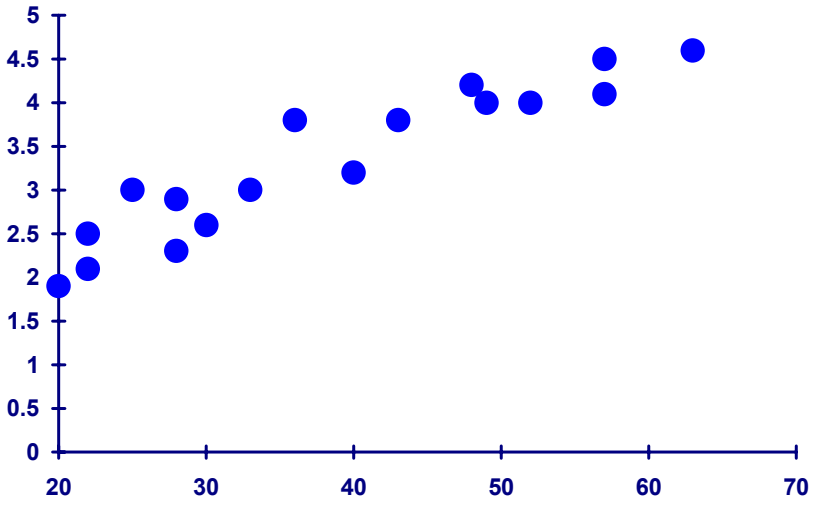
II. SIMPLE LINEAR REGRESSION ANALYSIS

We are now extending the idea of correlation into a rather mechanical concept called regression analysis. Before doing so, let us briefly recall this idea in the historical context. As mentioned earlier, In 1885, Francis Galton introduced the concept of "regression" in a study that demonstrated that offspring do not tend toward the size of parents, but rather toward the average as compared to the parents. The method of regression has, however, a longer history. In fact, a legendary French mathematician by the name of Adrien Marie Legendre published the first work on regression (although he did not use the word) in 1805. Still, the credit for discovery of the method of least squares generally given to Carl Friedrich Gauss (another legendary mathematician), who used the procedure in the early part of the 19th century.

Much used (and perhaps overused) cliche in data analysis "garbage in - garbage out" and "the results are only as good as the data that produced them" apply in the building of regression models. If the data do not reflect a trend involving the variables, there will be no success in model development or in drawing inferences regarding the system. Even with some types of relationship does exist, this does not imply that the data will reveal it in a clearly detectable fashion.

Many of the ideas and principles used in fitting linear models to data are best illustrated by using simple linear regression. These ideas can be extended to more complex modelling techniques once the basic concepts necessary for model development, fitting and assessment have been discussed.

Example 1 (continued): The plot of cholesterol (y-axis) versus age (x-axis) yields the following relationship:



From this graph, we can see that cholesterol level seems to vary systematically with age (which was confirmed earlier by the correlation analysis); moreover, the data points seem to scatter around the line connects between two points (20, 2.2) and (65, 4.5). Now, we learned earlier (in Topic 1) that for any two given points in a two-dimensional space, we could construct a straight line through two points. The same principle is applied here, although the technique of estimation is slightly more complicated.

2.1. ESTIMATES OF LINEAR REGRESSION MODEL

Let the observed pairs of values x and y be (x_1, y_1), (x_2, y_2), . . . , (x_n, y_n). The essence of a regression analysis is concerned with relationships between a response or dependent variable (y) and explanatory or independent variable (x). The simplest relationship is the straight line model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad [8]$$

In this model, β_0 and β_1 are unknown parameters and are to be estimated from the observed data, ε is a random error or departure term representing the level of inconsistency present in repeated observations under similar experimental conditions. To proceed with the parameter estimation, we have to make some assumptions

(i) The value of x is fixed (not random);

and on the random error ε, we assume that ε's are:

(i) normally distributed;
(ii) has expected value 0 i.e. E(ε) = 0
(iii) constant variance σ^2 for all levels of X;
(iv) and successively uncorrelated (statistically independent).

Because β_0 and β_1 are parameters (hence, constants) and that the value of x is fixed, we can obtain the expected value of [8] as :

$$E(y_i) = \beta_0 + \beta_1 x_i \qquad [9]$$

and
$$\begin{aligned} \text{var}(y_i) &= \text{var}(\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \text{var}(\varepsilon_i) \\ &= \sigma^2. \end{aligned} \qquad [10]$$

LEAST SQUARE ESTIMATORS

To estimate β_0 and β_1 from a series of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we use the method of least squares. This method estimates two constants b_0 and b_1 (corresponding to β_0 and β_1) so that they minimise the quantity:

$$Q = \sum_{i=1}^{n} [y_i - (b_0 + b_x x_i)]^2$$

It turns out that to minimise this quantity, we need to solve a system of simultaneous equations:

$$\sum y_i = nb_0 + b_1 \sum x_i$$

$$\sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2$$

And the estimates turn out to be:

$$\boxed{b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} = \frac{\text{cov}(x, y)}{\text{var}(x)} \qquad [11]$$

and $\boxed{b_0 = \bar{y} - b_1 \bar{x}}$ [12]

Example 1 (continued):

In our example, the estimates are:

$$b_1 = \frac{Cov(x, y)}{\text{cov}(x)} = \frac{10.68}{184.85} = 0.0577$$

and $b_0 = \bar{y} - b_1 \bar{x} = 3.33 - 0.0577(38.83) = 1.089.$

Hence the regression equation is:

$$y = 1.089 + 0.057x$$

That is, for any individual, his/her cholesterol is completely determined by the equation:

Cholesterol = 1.089 + 0.057(Age) + *e*

where *e* is the specific error which is not accounted for by the equation (including measurement error) associated with the subject. For instance, for subject 1 (46 years old), his/her expected cholesterol is: 1.089 + 0.057 x 46 = 3.7475; when compared with his/her actual value of 3.5, the residual is *e* = 3.5 - 3.7475 = -0.2475. Similarly, the expected cholesterol value for subject 2 is 1.089 + 0.057 x 26 = 2.245 and is higher than his/her actual level by 0.3450.

The predicted value calculated using the above equation, together with the residuals (*e*) are tabulated in the following table.

I.D	Observed (O)	Predicted (P)	Residual (e = O - E)
1	3.50	3.7475	-0.2475
2	1.90	2.2450	-0.3450
3	4.00	4.0942	-0.0942
4	2.60	2.8229	-0.2229
5	4.50	4.3832	0.1168
6	3.00	2.5339	0.4661
7	2.90	2.7073	0.1927
8	3.80	3.1696	0.6304
9	2.10	2.3606	-0.2606
10	3.80	3.5741	0.2259
11	4.10	4.3832	-0.2832
12	3.00	2.9962	0.00377
13	2.50	2.3606	0.1394
14	4.60	4.7299	-0.1299
15	3.20	3.4008	-0.2008
16	4.20	3.8631	0.3369
17	2.30	2.7073	-0.4073
18	4.00	3.9208	0.0792

2.2. TEST OF HYPOTHESIS CONCERNING REGRESSION PARAMETERS.

To some large extent, the interest will lie in the values of slope. Interpretation of this parameter is meaningless without a knowledge of its distribution. Therefore, having calculate the estimates b_1 and b_0, we need to determine the standard error of these parameters so that we can make inferences regarding their significance in the model. Before doing this, let us have a brief look at the significance of the term e.

We learned in earlier topic that if $\bar{y}$ is the sample mean of a variable Y, then the variance of Y is given by $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. Now, in the regression case, $\bar{y}$ is actually equal to $\beta_0 + \beta_1 x_i = \hat{y}$. Hence, it is reasonable that the sample variance of the residuals e

should provide an estimator of σ^2 in [10]. It is from this reasoning that the unbiased estimate of σ^2 is defined as:

$$\boxed{s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\left(e_i^2\right)} \qquad [13]$$

It can be shown that the expected values of b_1 and b_0 are β_1 and β_0 (true parameters), respectively. Furthermore, from [13], it can be shown that the variances of b_1 and b_0 are:

$$\boxed{\text{var}(b_1) = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad [14]$$

and $$\text{var}(b_0) = \text{var}(\bar{y}) + (\bar{x})^2\,\text{var}(b_1)$$

which is: $$\boxed{\text{var}(b_0) = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad [15]$$

Once can go a step further by estimating the covariance of b_1 and b_0 by:

$$\boxed{Cov(b_1, b_0) = -(\bar{x})^2 s^2(b_1)} \qquad [16]$$

That is, b_1 is normally distributed with mean β_1 and variance given in [14], and b_0 is normally distributed with mean β_0 and variance given in [15]. It follows that the test for significance of b_1 is the ratio

$$t = \frac{b_1}{\sqrt{\frac{s^2}{s_x^2}}} = \frac{b_1 s_x}{s}$$

which is distributed according to the *t* distribution with *n*-2 df.

and $$t = \frac{b_0}{s\sqrt{(1/n) + (\bar{x}^2 / s_x^2)}}$$

is a test for b_0, which is distributed according to the t distribution with n-2 df.

Example 1 (continued):

In our example, the estimate residual variance s^2 is calculated as follows:

$$s^2 = [(-0.2475)^2 + (-0.3450)^2 + \ldots + (0.0792)^2] / (18\text{-}2)$$
$$= 0.0916$$

We can calculate the corrected sum of square of AGE, $\sum_{i=1}^{n}(x_i - \bar{x})^2$, by working out from the estimate variance as:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2 \ (n - 1)$$
$$= 184.85 \ (17)$$
$$= 3142.45$$

Hence, the estimated variance of b_1 is:

$$\text{var}(b_1) = 0.0916 / 3142.45 = 0.00002914$$

$$\text{SE}(b_1) = \sqrt{\text{var}(b_1)} = 0.00539.$$

A test of hypothesis of $\beta_1 = 0$ can be constructed as:

$$t = b_1 / \text{SE}(b_1)$$
$$= 0.0578 / 0.00539$$
$$= 10.70$$

which is highly significant ($p < 0.0001$).

For the intercept we can estimate its variance as:

$$\text{var}(b_0) = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$
$$= 0.0916\left(\frac{1}{19} + \frac{(38.83)^2}{3142.45}\right)$$
$$= 0.049$$

And the test of hypothesis of $\beta_0 = 0$ can be constructed as:

$$t = b_0 \,/\, \text{SE}(b_0)$$
$$= 1.089 \,/\, \sqrt{0.049}$$
$$= 4.92$$

which is also highly significant ($p < 0.001$).

2.3. ANALYSIS OF VARIANCE

An analysis of variance partitions the overall variation between the observations Y into variation which has been accounted for by the regression on X and residual or unexplained variation. Thus, we can say:

Total variation about the mean	=	Variation explained by regression model	+	Residual variation

In ANOVA notation, we can write equivalently:

SSTO = SSR + SSE

or,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Now, SSTO is associated with n-1 df. For SSR, there are two parameters (b_0 and b_1) in the model, but the constraint $\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = 0$ takes away 1df, hence it has finally 1 df. For SSE, there are n residuals (e_i); however, 2 df are lost because of two constraints on the e_i's associated with estimating the parameters β_0 and β_1 by the two normal equations see section 2.1).

We can assemble these data in an ANOVA table as follows:

Source	df	SS	MS
Regression	1	SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MSR = SSR/1
Residual error	n - 2	SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	MSE = SSE / (n-2)
Total	n - 1	SSTO = $\sum_{i=1}^{n}(y_i - \bar{y})^2$	

R-SQUARE

From this table it seems to be sensible to obtain a "global" statistic to indicate how well the model fits the data. If we divide the regression sum of square (variation due to regression model, SSR) by the total variation of Y (SSTO), we would have what statisticians called the **coefficient of determination**, which is denoted by R^2:

$$: R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad [17]$$

In fact, it can be shown that the coefficient of correlation r defined in [5] is equal to $\sqrt{R^2}$.

Obviously, R^2 is restricted to $0 \leq R^2 \leq 1$. An $R^2 = 0$ indicates that *X* and *Y* are independent (unrelated), whereas an $R^2 = 1$ indicates that *Y* is completely determined by *X*. However, there are a lot of pitfalls in this statistic. A value of $R^2 = 0.75$ is likely to be viewed with some satisfaction by experimenters. It is often more appropriate to recognise that there is still another 25% of the total variation unexplained by the model. We must ask why this could be, and whether a more complex model and/or inclusion of additional independent variables could explain much of this apparently residual variation.

A large R^2 value does *not* necessarily mean a good model. Indeed, R^2 can artificially high when either the slope of the equation is large or the spread of the independent variable is large. Also a large R^2 can be obtained when straight lines are fitted to data that display non-linear relationships. Additional methods for assessing the fit of a model are therefore needed and will be described later.

F STATISTIC

An assessment of the significance of the regression (or a test of the hypothesis that $\beta_1 = 0$) is made from the ratio of the regression mean square (MSR) to the residual mean square MSE (s^2) which is an F-ratio with 1 and *n*-2 degrees of freedom. This calculation is usually exhibited in an analysis of variance table produced by most computer programs.

$$\boxed{F = \frac{MSR}{MSE}} \qquad [18]$$

It is important that a highly significant F ratio should not seduce the experimenter to a belief that the straight line fits the data superbly. The F test is simply an assessment of the extend to which the fitted line has a slope which is different from zero. If the slope of the line is near zero, the scatter of the data points about the line would need to be small in order to obtain a significant F ratio. However, a situation with a slope very different from zero can give a highly significant F ratio with a considerable scatter of points about the line.

The F test as defined in [18] is actually equivalent to the *t* test in $t = \dfrac{b_1}{\sqrt{\dfrac{s^2}{s_x^2}}} = \dfrac{b_1 s_x}{s}$.

The F test is therefore can be used for testing $\beta_1 = 0$ versus $\beta_1 \neq 0$ and is not for testing one-sided alternatives.

Example 1 (continued):

In our example, the sum of squares due to regression line is:

$$\begin{aligned} \text{SSR} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ &= (3.7475 - 3.33)^2 + (2.2450 - 3.33)^2 + \ldots + (3.9208 - 3.33)^2 \\ &= 10.4944 \end{aligned}$$

which is associated with 1 df, hence its mean square is 10.4944.

The sum of squares due to residuals is:

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= (-0.2475)^2 + (0.3450)^2 + \ldots + (0.0792)^2 \\ &= 1.4656 \end{aligned}$$

and is associated with 18-2 = 16 df, hence its mean square is 1.4656 / 16 = 0.0916

The F statistic is then:

$$F = 10.4944 / 0.0916 = 114.565.$$

Hence, the ANOVA table can be set up as follows:

Source	df	SS	MS	F-test

Regression	1	10.4944	10.4944	114.565
Residual errors	16	1.4656	0.0916	
Total	17	11.960		

Accordingly, the coefficient of determination is: $R^2 = 10.49 / 11.96 = 0.8775$. This means that 87.75% of total variation in cholesterol between subjects is "explained" by the regression equation. //

2.3. Analysis of Residuals And The Diagnosis of Regression Model

A residual is defined as the difference between the observed and predicted y value, given by $e_i = y_i - \hat{y}_i$, the value which is not accounted for by the regression equation. Hence, an examination of this term should reveal how appropriate the equation is.

However, these residuals do not have constant variance. In fact, var(e_i) = (1-h_i)s^2, where h_i is the ith diagonal element of the matrix H which is such that $\hat{y}_i$ = Hy. H is called the "hat matrix", since it defines the transformation that puts the "hat" on y! In view of this, it is preferable to work with the standardised residuals. In simple linear regression case, the standardised residual r_i is defined as:

$$r_i = \frac{e_i}{\sqrt{MSE}} \qquad [19]$$

These standardised residuals have mean 0 and variance 1. We can use r_i to verify assumptions of the regression model which we made in section 2.1. These are:

(a) are the regression function is not linear;

(b) the distributions of Y (cholesterol) do not have constant variance at all level of X (age) or equivalently the residuals do not have constant variances;

(c) the distributions of Y are not normal or equivalently the residuals are not normal;

(d) the residuals are not independent.

Useful graphical methods for examining the standard assumptions of constant variance, normality of the error terms and appropriateness of the fitted model include:

- A plot of residuals against fitted values to identify outliers, detect systematic departure from the model or detect non-constant variance;

- A plot of residuals in the order in which the observations were taken to detect non-independence.

- A normal probability plot of the residuals to detect from normality.

- A plot of residuals against X can indicate whether a transformation of the original X variable is necessary, while a plot of residuals against X variables omitted from the model could reveal whether the y variable depends on the omitted factors.

OUTLIERS

Outliers in regression are observations that are not well fitted by the assumed model. Such observations will have large residuals. A crude rule of thumb is that an observation with a standardised residual greater than 2.5 in absolute value is an outlier and the source of that data point should be investigated, if possible. More often than not, the only evidence that something has gone wrong in the data generating process is provided by the outliers themselves ! A sensible way of proceeding with the analysis is to determine whether those values have substantial effect on the inferences to be drawn from the regression analysis, that is, whether they are influential.

INFLUENTIAL OBSERVATIONS

Generally speaking, it is more important to focus on influential outliers. But it is not only outliers that can be influential. If observation is separated from the others in terms of the values of the X-variables, this observation is likely to influence the fitted regression mode. Observations separated from other in this way will have a large value of h_i. We call h_i the leverage, A rough guide is that observations with $h_i > 3p/n$ are influential, where p is the number of beta coefficients in the model (in our example $p = 2$).

There is a problem (drawback) to using the leverage to identify influential values - it does not contain any information about the value of the *Y* variable, only the value of the *X* variables. To detect an influential observation, a natural statistic to use is a scaled version of $\left(\hat{y}_{i(j)} - y_i\right)^2$ where $\hat{y}_{i(j)}$ is the fitted value for the *j*th observation when the *i*th observation is omitted from the fit. This leads to the so-called Cook's statistic. Fortunately, to obtain the value of this statistic, we do not need to carry out a regression fit, omitting each point in term, for the statistic given by:

$$D_i = \frac{r_i^2 - h_i}{p(1 - h_i)}$$

Observations with relatively large values of D_i are defined as influential.

Example 1 (continued): Calculations of studentised residuals and Cook's D statistic for each observation are given in the following table:

ID	Observed	Predicted	Std Err	Std. Res			Cook's D
1	3.5000	3.7475	0.292	-0.849	*|	|	0.028
2	1.9000	2.2450	0.276	-1.250	**|	|	0.158
3	4.0000	4.0942	0.285	-0.330	|	|	0.007
4	2.6000	2.8229	0.290	-0.768	*|	|	0.026
5	4.5000	4.3832	0.277	0.421	|	|	0.017
6	3.0000	2.5339	0.284	1.638	|***	|	0.177
7	2.9000	2.7073	0.288	0.669	|*	|	0.023
8	3.8000	3.1696	0.294	2.146	|****	|	0.142
9	2.1000	2.3606	0.280	-0.931	*|	|	0.074
10	3.8000	3.5741	0.293	0.770	|*	|	0.019
11	4.1000	4.3832	0.277	-1.021	**|	|	0.100
12	3.0000	2.9962	0.292	0.013	|	|	0.000
13	2.5000	2.3606	0.280	0.498	|	|	0.021
14	4.6000	4.7299	0.264	-0.493	|	|	0.039
15	3.2000	3.4008	0.294	-0.683	*|	|	0.014
16	4.2000	3.8631	0.290	1.162	|**	|	0.061
17	2.3000	2.7073	0.288	-1.413	**|	|	0.102
18	4.0000	3.9208	0.289	0.274	|	|	0.004

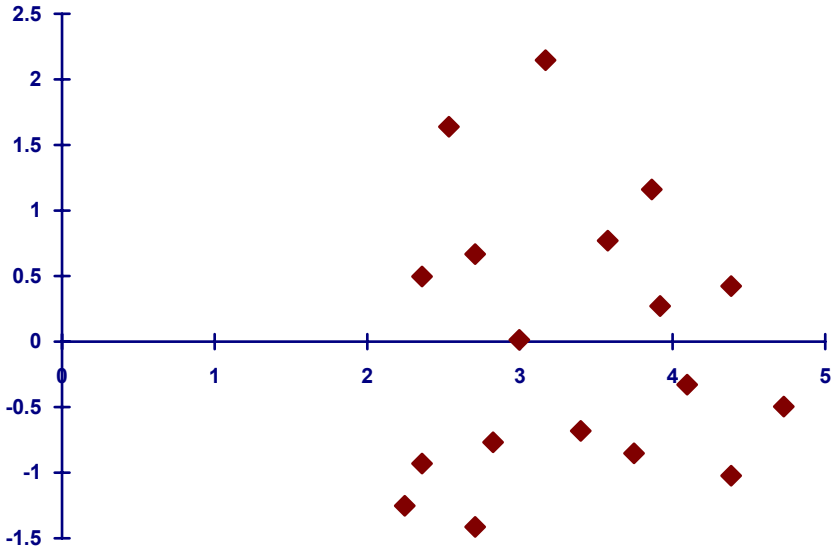


Figure 1: Plot of standardised residuals against predicted value of *y*.

2.4. SOME FINAL COMMENTS

(A) INTERPRETATION OF CORRELATION

The following is an extract from D Altman's comments:

"Correlation coefficients lie within the ranged -1 to +1, with the midpoint of zero indicating no linear association between the two variables. A very small correlation does not necessarily indicate that two variables are not associated, however. To be sure of this, we should study a plot of the data, because it is possible that the two variables display a peculiar (i.e. non-linear) relationship. For example, we should not observe much, if any, correlation between the average midday temperature and calendar moth because there is a cyclic pattern. More common is the situation of a curved relationship between two variables, such as between birthweight and length of gestation. In this case, Pearson's *r* will underestimate the association as it is a measure of linear association. The rank correlation coefficient is better here as it assesses in a more general way whether the variables tend to rise together (or move in opposite direction).

It is surprising how unimpressive a correlation of 0.5 or even 0.7 is when a correlation of this magnitude is significant at $p<0.05$ level with a sample size of 9 or 15 subjects. Whether these are important is another matter. Feinstein commented on the lack of clinical relevance of a statistically significant correlation of less than 0.1 found in a sample of 6000. The problem of clinical relevance is one that must be judged on its merits in each case, and depends on the context. For example, the same small correlation may be important in an epidemiological study but unimportant clinically.

One way of looking at the correlation that helps to modify the over-enthusiasm is to calculate the R-square value, which is the percentage of the variability of the data that is "explained" by the association between the two variables. So, a correlation of 0.7 implies that just 49% of the variability may be put down to the observed association.

Interpretation of association is often problematic because *causation can not be directly inferred*. If we observe an association between two variables X and Y, there are several possible explanations. Excluding the possibility that it is a chance finding, it may be because:

X causes (influences) Y
Y influences X
Both X and Y are influenced by one or more other variables.

Correlation is often used as an exploratory method for investigating inter-relationships among many variables, for which purpose it is most obvious to use hypothesis tests. Although in principle, this approach is often over-done. The problem is that even with a modest number of variables, the number of coefficients is large: 10 variables yield (10 x 9)/2 = 45 *r* values, and 20 variables yields 190 *r* values. On of the 20 of these will be significant at the level of 5% purely by chance, and these can not be distinguished from genuine association. Furthermore, the magnitude of correlation that is significant at 5% is dependent on sample size. In a large sample,, even if there are several significant *r* values of around 0.2 to 0.3, say, these are unlikely to be very useful. While this way of looking at large numbers of variables can be helpful when one really has no prior hypothesis, significant association really needs to be confirmed in another set of data before credence can be given to them.

Another common problem of interpretation occurs when we know that each of two variables is associated with a third variables. For example, if X is positively correlated with Y and Y is positively correlated with Z, it is tempting to say that X and Z must be positively correlated. Although this may indeed be true, such an inference is unjustified - we can not say anything about the correlation between X and Z. The same is true when one has observed no association. For example, Mazess et al (1984) the correlation between age and height was 0.05 and between weight and %fat was 0.03. This does not imply that the correlation between age and %fat was also near zero. In fact, this correlation was 0.79. *Correlation can not be inferred from direct associations*."

(B) INTERPRETATION OF REGRESSION

The variability among a set of observations may be partly attributed to known factors and partly to unknown factors; the latter is often termed "random variation". In linear regression, we see how much of the variability in the response variable can be attributed to different values of the predictor variable, and the scatter either side of the fitted line shows unexplained variability. Because of this variability, the fitted line is only an estimated of the relation between these variables in the population. As with other

estimates (such as a sample mean) there will be uncertainty associated with the estimated slope and intercept. The confidence intervals for the whole line and prediction intervals for individual subjects show other aspect of variability. The latter are especially useful as regression is often used to make predictions about individuals.

It should be remembered that the regression line should not be used to make predictions for X values outside the range of values in the observed data. Such extrapolation is unjustified as we have no evidence about the relationship beyond the observed data. A statistical model is only an approximation. One rarely believes, for example, that the true relationship is exactly linear, but the linear regression equation is taken as a reasonable approximation for the observed data. Outside the range of the observed data one can not safely use the same equation. Thus, we should not use the regression equation to predict value beyond what we have observed.

III. EXERCISES

1. Consider the simple regression equation [8]. Consider the least square residuals, given by $y_i - \hat{y}_i$ where $i = 1, 2, 3, \ldots, n$. Show that

   (a) $\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}$. (b) $\sum_{i=1}^{n} (\hat{y}_i - \bar{y}) = 0$

2. The following data represent diastolic blood pressures taken during rest. The x values denote the length of time in minutes since rest began, and the y values denote diastolic blood pressures.

x:	0	5	10	15	20
y:	72	66	70	64	66

   (a) Construct a scatter plot.
   (b) Find the coefficient of correlation and the linear regression equation of y on x.
   (c) Calculate 95% confidence interval for the slope and intercept.
   (d) Calculate 95% confidence interval for the predicted value of y when $x = 10$.

3. The following table shows resting metabolic rate (RMR) (kcal/24 hr) and body weight (kg) of 44 women (Owen et al 1986).

Wt:	49.9	50.8	51.8	52.6	57.6	61.4	62.3	64.9	43.1	48.1	52.2
RMR:	1079	1146	1115	1161	1325	1351	1402	1365	870	1372	1132
Wt:	53.5	55.0	55.0	56.0	57.8	59.0	59.0	59.2	59.5	60.0	62.1
RMR:	1172	1034	1155	1392	1090	982	1178	1342	1027	1316	1574
Wt:	64.9	66.0	66.4	72.8	74.8	77.1	82.0	82.0	83.4	86.2	88.6
RMR:	1526	1268	1205	1382	1273	1439	1536	1151	1248	1466	1323
Wt:	89.3	91.6	99.8	103	104.5	107.7	110.2	122.0	123.1	125.2	143.3
RMR:	1300	1519	1639	1382	1414	1473	2074	1777	1640	1630	1708.

   (a) Perform linear regression analysis of RMR on body weight.
   (b) Examine the distribution of residuals. Is the analysis valid?

(c) Obtain a 95% confidence interval for the slope of the line.
(d) Is it possible to use an individual's weight to predict their RMR to within 250 kcal/24 hr ?

4. Digoxin is a drug that is largely eliminated unchanged in the urine. In a study by Halkin et al. 1975, in which the authors commented:
(a) Its renal clearance was correlated with creatinine clearance;
(b) it clearance was independent of urine flow.
The authors provided data showing measurements of these three variables from 35 inpatients being treated with digoxin for congestive heart failure.

Patient	Clearance Creatinine (ml/min/1.73 m^2)	Digoxin	Urin flow (ml/min)
1	19.5	17.5	0.74
2	24.7	34.8	0.43
3	26.5	11.4	0.11
4	31.1	29.3	1.48
5	31.3	13.9	0.97
6	31.8	31.6	1.12
7	34.1	20.7	1.77
8	36.6	34.1	0.70
9	42.4	25.0	0.93
10	42.8	47.4	2.50
11	44.2	31.8	0.89
12	49.7	36.1	0.52
13	51.3	22.7	0.33
14	55.0	30.7	0.80
15	55.9	42.5	1.02
16	61.2	42.4	0.56
17	63.1	61.1	0.93
18	63.7	38.2	0.44
19	66.8	37.5	0.50
20	72.4	50.1	0.97
21	80.9	50.2	1.02
22	82.0	50.0	0.95

23	82.7	31.8	0.76
24	87.9	55.4	1.06
25	101.5	110.6	1.38
26	105.0	114.4	1.85
27	110.5	69.3	2.25
28	114.2	84.8	1.76
29	117.8	63.9	1.60
30	122.6	76.1	0.88
31	127.9	112.8	1.70
32	135.6	82.2	0.98
33	136.0	46.8	0.94
34	153.5	137.7	1.76
35	201.1	76.1	0.87

(a) Do these data support statements (a) and (b) above?

5. Consider the following 4 data sets. Note that $X1 = X2 = X4$. Fit a linear regression equation with Y as a dependent variable and X as an independent variable. What is the most striking feature from these data sets. Carry out a residual plot for each data set and comment on the result.

$X1$	$Y1$	$X2$	$Y2$	$X3$	$Y3$	$X4$	$Y4$
4	4.26	4	3.1	8	6.58	4	5.39
5	5.68	5	4.74	8	5.76	5	5.73
6	7.24	6	6.13	8	7.71	6	6.08
7	4.82	7	7.26	8	8.84	7	6.42
8	6.95	8	8.14	8	8.47	8	6.77
9	8.81	9	8.77	8	7.04	9	7.11
10	8.04	10	9.14	8	5.25	10	7.46
11	8.33	11	9.26	8	5.56	11	7.81
12	10.84	12	9.13	8	7.91	12	8.15
13	7.58	13	8.74	8	6.89	13	12.74
14	9.96	14	8.10	9	12.50	14	8.84