ANALYSIS OF COVARIANCE

Blocking Factor is basically a nuisance factor in the experiment which has effect on the response but its effect is not of interest to experimenter. We can think of three situations as below:

1. It is unknown or not measurable, we can try to randomize the experiment to average out its impact.
2. If known and controllable, we can use blocking to block its effect and it becomes a blocking factor.
3. If known and measurable but uncontrollable, then we can use Analysis of covariance.

An Example:

An experiment was designed to study the performance of four detergents in cleaning clothes. The following cleanliness readings obtained with specifically designed equipment for three types of common stains. Is there a difference between the detergents?

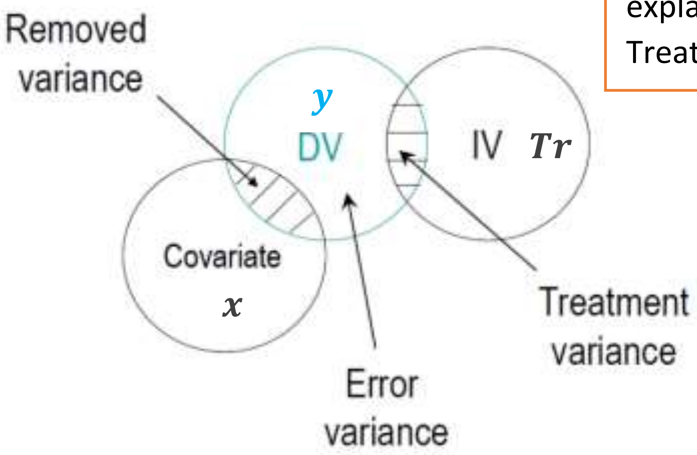
	Stain 1	Stain 2	Stain 3
Detergent 1	45	43	51
Detergent 2	47	46	52
Detergent 3	48	50	55
Detergent 4	42	37	49

In above example, we are interested in the effect of detergents on cleanliness, three type of stains may have effect on the response but we are not interested in it. So, stain is a block factor and by blocking effect of Stain Type, we are basically improving the precision with which comparisons between Detergent Type effects are made.

The analysis of covariance is another technique that is occasionally useful for improving the precision of an experiment. Suppose that in an experiment with a response variable *y*, there is another variable, say *x*, and that y is linearly related to x. Furthermore, suppose that x cannot be controlled by the experimenter but can be observed along with y. The variable x is called a **covariate** or **concomitant variable**. The analysis of covariance involves adjusting the observed response variable for the effect of the concomitant variable. If such an adjustment is not performed, the concomitant variable could inflate the error mean square and make true differences in the response due to treatments harder to detect. Thus, the analysis of covariance is a method of adjusting for the effects of an uncontrollable nuisance variable. As we will see, the procedure is a combination of analysis of variance and regression analysis.

Treatment variation is the portion of variation in Dependent Variable (y) explained by the Treatment Variable (Tr).

Graphical Representation:

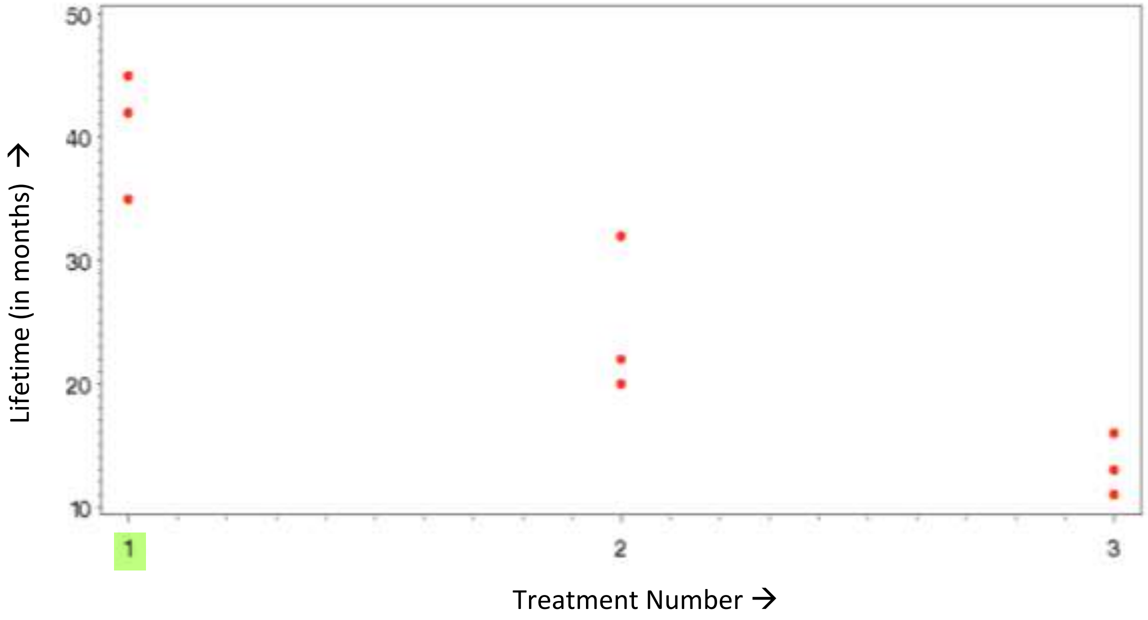


The covariate removes certain variation in the dependent variable, thereby reducing the error variation.

In some situations, failure to include an important covariate can yield misleading results.

Let us take an example of studying the possible impact of three Treatments on an aggressive form of cancer. Response variable is the number of months a patient survives after being subjected to a treatment.

Plot of the data on lifetime (in months) versus Treatments are as below:

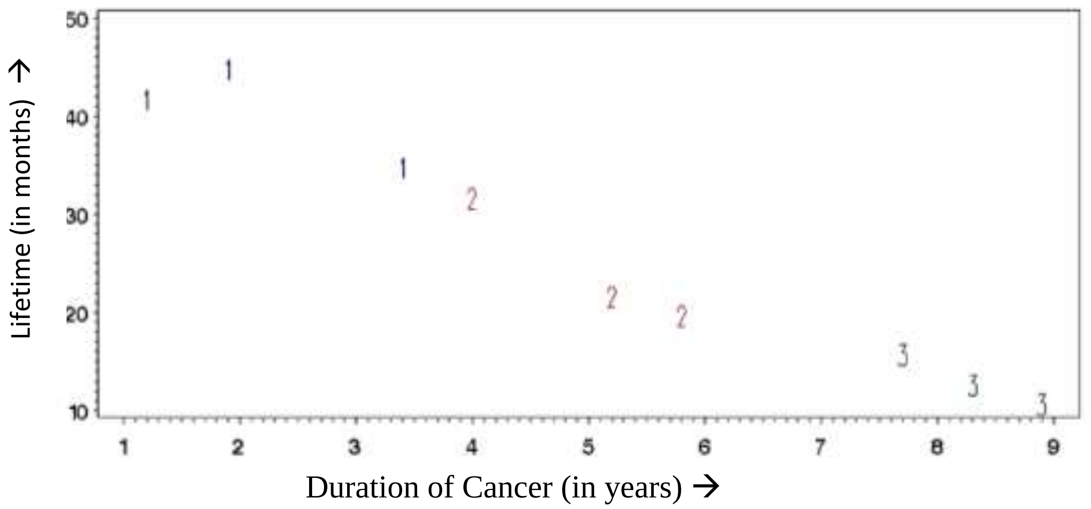


From above graph, it appears that treatments has significant impact on lifetimes and Treatment 1 is clearly the best, since people live longer.

So, the researcher put a large group of people on Treatment 1 expecting them to survive 40+ months, but unfortunately they do not live that long. It means that something is missing somewhere.

So, it was thought to consider the stage to which the cancer has progressed at the time of administering the Treatment. This was important because those at earlier stages of the disease will naturally live longer on the average.

The following plot of lifetime (in moths) versus duration of cancer (in years) for different Treatments illustrates where things went wrong.



There is clearly a linear relationship between the duration of cancer and the length of time the respective person survived.

Furthermore, we notice from the plot that the group assigned to the first treatment were all in their earlier stage of the disease, those assigned to the second treatment were all in a middle stage, and those assigned to the third treatment were all in a later stage.

After seeing this plot, it is clear that we can’t or shouldn’t compare the lifetimes without considering the duration of the disease. So, one would suspect, by looking at this plot, that the treatments may not have significant impact on Lifetime.

So, in such a situation, one must use ANCOVA to arrive at correct decision by removing the effect of the concomitant variable.

Another Example

As an example, consider a study performed to determine if there is a difference in the strength of certain type of fibre produced by three machines. However, the strength of the fibre is also affected by its thickness and, consequently, a thicker fibre will generally be stronger than a thinner one. Data, on breaking strength and fibre diameter, collected using a random sample of five fibre specimens from each machine is given below:

Table 1: Breaking Strength Data
(y = strength in pounds and x = diameter in 10^{-3} inches)

Machine 1		Machine 2		Machine 3	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15

Figure below presents a scatter diagram of strength (y) versus the diameter (or thickness) of the sample. Clearly the strength of the fibre is also affected by its thickness; consequently, a thicker fibre will generally be stronger than a thinner one. The analysis of covariance could be used to remove the effect of thickness (x) on strength (y) when testing for differences in strength between machines.

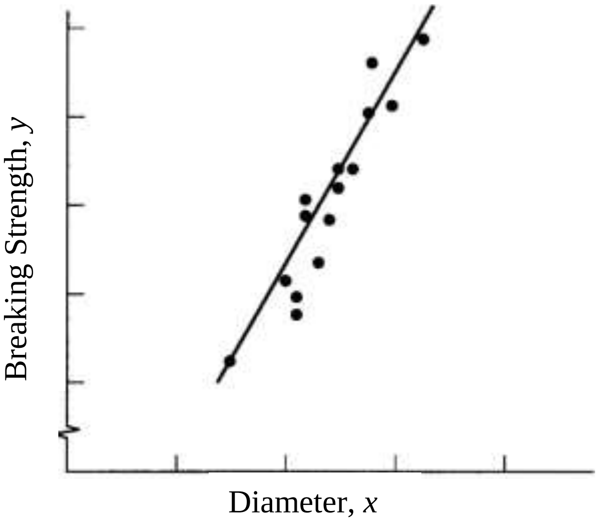


Figure 1: Scatter Diagram between diameter and breaking strength.

Description of the Procedure

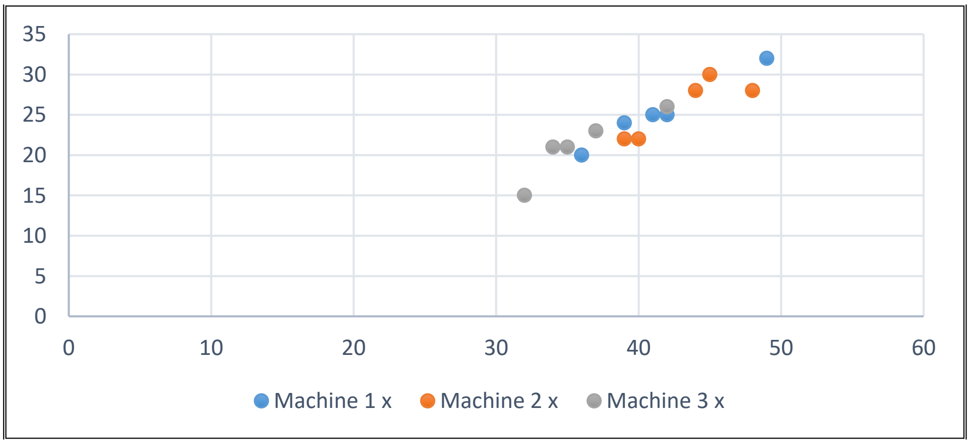
The basic procedure for the analysis of covariance is described and illustrated for a single-factor experiment with one covariate. Assuming that there is a linear relationship between the response and the covariate, an appropriate statistical model is

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \bar{x}_{\circ\circ}) + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,n \end{cases} \qquad (1)$$

where y_{ij} is the *j*-th observation on the response variable taken under the *i*th treatment or level of the single factor, x_{ij}, is the measurement made on the covariate or concomitant variable corresponding to y_{ij}; (i.e., the *ij*-th run), $\bar{x}_{..}$, is the mean of the x_{ij} values, μ is an overall mean of *y*-values, τ_i is the effect of the *i*th treatment, β is a linear regression coefficient indicating the dependency of y_{ij} on x_{ij}, and ε_{ij}, is a random error component. We assume

a. the errors $\varepsilon_{ij}$are $NID(0, \sigma^2)$,
b. the slope $\beta \neq 0$ and true relationship between y_{ij} and x_{ij} is linear,
c. the regression coefficients for each treatment are nearly identical, in other words regression lines are parallel,
d. the treatment effects sum to zero ($\sum_{i=1}^{a} \tau_i = 0$), and
e. the concomitant variable x_{ij} is not affected by the treatments.

This model assumes that all treatment regression lines have identical slopes. If the treatment interacts with the covariates, the slopes may be non-identical. Covariance analysis is not appropriate in such cases.



It may be noted from Equation (1) that the analysis of covariance model is a combination of the linear models employed in analysis of variance and regression. That is, we have treatment effects $\{\tau_i\}$ as in a single-factor analysis of variance and a regression coefficient β as in a regression equation. The concomitant variable in Equation (1) is expressed as $(x_{ij} - \bar{x}_{..})$ instead of x_{ij} so that the parameter μ is preserved as the overall mean. The model could have been written as

$$y_{ij} = \mu' + \tau_i + \beta x_{ij} + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,n \end{cases} \qquad (2)$$

where μ' is a constant not equal to the overall mean, which for this model is $\mu - \beta\bar{x}_{..}$. Equation (1) is more widely found in the literature.

To describe the analysis, we introduce the following notations.

$$S_{yy} = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})y_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{an}$$

$$S_{xx} = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})x_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}^2 - \frac{x_{..}^2}{an}$$

$$S_{xy} = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..}) = \sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{..})y_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}y_{ij} - \frac{(x_{..})(y_{..})}{an}$$

$$T_{yy} = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 = n\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})\bar{y}_{i.} = \sum_{i=1}^{a}\frac{y_{i.}^2}{n} - \frac{y_{..}^2}{an}$$

$$T_{xx} = n\sum_{i=1}^{a}(\bar{x}_{i.} - \bar{x}_{..})^2 = n\sum_{i=1}^{a}(\bar{x}_{i.} - \bar{x}_{..})\bar{x}_{i.} = \sum_{i=1}^{a}\frac{x_{i.}^2}{n} - \frac{x_{..}^2}{an}$$

$$T_{xy} = n\sum_{i=1}^{a}(\bar{x}_{i.} - \bar{x}_{..})(\bar{y}_{i.} - \bar{y}_{..}) = n\sum_{i=1}^{a}(\bar{x}_{i.} - \bar{x}_{..})\bar{y}_{i.} = \sum_{i=1}^{a}\frac{(x_{i.})(y_{i.})}{n} - \frac{(x_{..})(y_{..})}{an}$$

$$E_{yy} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot}\right)y_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{a}\frac{y_{i\cdot}^2}{n} = S_{yy} - T_{yy}$$

$$E_{xx} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\cdot}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\cdot}\right)x_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}^2 - \sum_{i=1}^{a}\frac{x_{i\cdot}^2}{n} = S_{xx} - T_{xx}$$

$$E_{xy} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\cdot}\right)\left(y_{ij} - \bar{y}_{i\cdot}\right) = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\cdot}\right)y_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{n} x_{ij}y_{ij} - \sum_{i=1}^{a}\frac{(x_{i\cdot})(y_{i\cdot})}{n} = S_{xy} - T_{xy}$$

Note that, in general, $S = T + E$, where the symbols S, T, and E are used to denote sums of squares and cross-products for total, treatments, and error respectively. The sums of squares for x and y must be nonnegative; however, the sums of cross-products (xy) may be negative.

Development by the General Regression Significance Test

It is possible to develop formally the procedure for testing $H_0 : \tau_i = 0$ in the covariance model

$$y_{ij} = \mu + \tau_i + \beta\left(x_{ij} - \bar{x}_{\circ\circ}\right) + \varepsilon_{ij} \qquad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,n \end{cases}$$

using the general regression significance test.

We can estimate the parameters of the above model by method of least squares. The least squares function is

$$L = \sum_{i=1}^{a}\sum_{j=1}^{n}\left[y_{ij} - \mu - \tau_i - \beta\left(x_{ij} - \bar{x}_{\bullet\bullet}\right)\right]^2$$

and from $\partial L/\partial\mu = \partial L/\partial\tau_i = \partial L/\partial\beta = 0$, we obtain the following $(a + 2)$ normal equations

$$\mu: \quad an\hat{\mu} + n\sum_{i=1}^{a}\hat{\tau}_i = y_{\bullet\bullet} \qquad (3)$$

$$\tau_i \ : n\hat{\mu} + n\hat{\tau}_i + \hat{\beta}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{\cdot\cdot}\right) = y_{i\cdot} \qquad i = 1,2,\cdots,a \qquad (4)$$

$$\beta: \sum_i \sum_j [y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}(x_{ij} - \bar{x}_{..})](x_{ij} - \bar{x}_{..}) = 0$$

$$\Rightarrow \sum_i \hat{\tau}_i \sum_j (x_{ij} - \bar{x}_{..}) + \hat{\beta} S_{xx} = \sum_i \sum_j (x_{ij} - \bar{x}_{..})(y_{ij} - \hat{\mu}) \qquad (5)$$

Since we have assumed that the treatments effects add to zero, that is $\sum_{i=1}^{a} \tau_i = 0$, we obtain from Equation 3,

$$\hat{\mu} = \bar{y}_{\bullet\bullet}$$

and from Equation 4

$$\hat{\tau}_i = (\bar{y}_{i.} - \bar{y}_{..}) - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..})$$

Equation 5, after substitution of $\hat{\tau}_i$ and $\hat{\mu}$, may be rewritten as

$$\sum_{i=1}^{a} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \sum_{j=1}^{n} (x_{ij} - \bar{x}_{\bullet\bullet}) - \hat{\beta} \sum_{i=1}^{a} (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) \sum_{j=1}^{n} (x_{ij} - \bar{x}_{\bullet\bullet}) + \hat{\beta} S_{xx} = S_{xy} \;.$$

But it can be shown that

$$\sum_{i=1}^{a} (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) \sum_{j=1}^{n} (x_{ij} - \bar{x}_{\bullet\bullet}) = T_{xy}$$

and

$$\sum_{i=1}^{a} (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) \sum_{j=1}^{n} (x_{ij} - \bar{x}_{\bullet\bullet}) = T_{xx} .$$

Therefore, the solution to Equation 5 is

$$\hat{\beta} = \frac{S_{xy} - T_{xy}}{S_{xx} - T_{xx}} = \frac{E_{xy}}{E_{xx}}$$

It may be noted that by fitting a model to the data set, we "**explain**" certain amount of the variation in the dependent variable, *y* and thereby obtain an estimate of the unexplained variation (or, SS_E) by subtracting the explained variation from total variation (i.e. SS_T). We can, in general, then have this expression $\boldsymbol{SS_E = SS_T - SS_{\text{Explained}}}$.

Referring to the expression $\boldsymbol{SS_E = y^T y - \hat{b}^T X^T y}$ and noting that normal equation is $\boldsymbol{X^T X b = X^T y}$, we see that the explained variation ($\hat{b}^T X^T y$) is always the sum of the parameter estimates, each multiplied by the right hand side of the corresponding normal equation. This is also known as reduction in sum of squares due to fitting a model.

Thus the reduction in the sum of squares due to fitting the full model involving $\mu, \{\tau_i\}$ and β can be expressed as

$$R(\mu, \tau, \beta) = \hat{b}^T \{X^T y\} = [\hat{\mu} \quad \{\hat{\tau}_i\} \quad \hat{\beta}] \begin{bmatrix} y_{..} \\ \{y_{i.}\} \\ S_{xy} \end{bmatrix}$$

i.e.

$$\begin{aligned} R(\mu, \tau, \beta) &= \hat{\mu} y_{\bullet\bullet} + \sum_{i=1}^{a} \hat{\tau}_i y_{i\bullet} + \hat{\beta} S_{xy} \\ &= (\bar{y}_{\bullet\bullet}) y_{\bullet\bullet} + \sum_{i=1}^{a} \left[(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}) - (E_{xy}/E_{xx})(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}) \right] y_{i\bullet} + (E_{xy}/E_{xx}) S_{xy} \\ &= y_{\bullet\bullet}^2/an + T_{yy} + (E_{xy}/E_{xx})(S_{xy} - T_{xy}) \\ &= y_{\bullet\bullet}^2/an + T_{yy} + (E_{xy})^2 / E_{xx} \end{aligned}$$

This Sum of squares has a + 1 degrees of freedom, since among the normal equations number of linear independent equations are a +1. Since $\sum_{i=1}^{a} \tau_i = 0$, number of independent τ_i's will be $(a - 1)$ and hence number of independent normal equations are $1 + (a - 1) + 1 = a + 1$.

The error sum of squares for the full model is thus

$$\begin{aligned} SS_E &= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - R(\mu, \tau, \beta) \\ &= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - y_{\bullet\bullet}^2 / an - T_{yy} - \left(E_{xy}\right)^2 / E_{xx} \\ &= S_{yy} - T_{yy} - \left(E_{xy}\right)^2 / E_{xx} \\ &= E_{yy} - \left(E_{xy}\right)^2 / E_{xx} \end{aligned}$$

with $an - (a+1) = a(n-1) - 1$ degrees of freedom.

It may be noted that SS_E, as obtained above, can be thought of as the error sum of squares for *y* adjusted due to regression of *y* on the concomitant variable *x*.

Now, consider the model restricted by the null hypothesis, that is, $H_0: \tau_i = 0$. Under this condition, the reduced model is

$$y_{ij} = \mu + \beta\left(x_{ij} - \bar{x}_{\circ\circ}\right) + \varepsilon_{ij} \quad \begin{cases} i = 1,2,\cdots,a \\ j = 1,2,\cdots,n \end{cases} \qquad (6)$$

Proceeding as earlier, the normal equations for this model are

$$\mu: \quad an\hat{\mu} = y_{\bullet\bullet}$$

$$\beta: \quad \hat{\beta} S_{xx} = S_{xy}$$

So, we get the estimates as

$$\hat{\mu} = \bar{y}_{\bullet\bullet} \quad \text{and} \quad \hat{\beta} = S_{xy} / S_{xx}$$

and the reduction in the sum of squares due to fitting of the reduced model is

$$R(\mu, \beta) = \hat{\mu} y_{\bullet\bullet} + \hat{\beta} S_{xy} \quad = y_{\bullet\bullet}^2 / an + \left(S_{xy}\right)^2 / S_{xx}$$

This sum of square has 2 degrees of freedom as both the normal equations are linearly independent in this case. The error sum of square for this model is

$$
\begin{aligned}
SS'_E &= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - R(\mu, \beta) \\
&= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - y_{\bullet\bullet}^2 / an - (S_{xy})^2 / S_{xx} \\
&= S_{yy} - (S_{xy})^2 / S_{xx}
\end{aligned}
$$

with *an* – 2 degrees of freedom.

Similarly, here it can be assumed that SS'_E is the total SS for *y* adjusted due to the regression of *y* on the concomitant variable *x*.

So, the sum of squares for testing the null hypothesis of no treatment effect, i.e. τ = 0 is

$$
\begin{aligned}
R(\tau|\mu, \beta) &= R(\mu, \tau, \beta) - R(\mu, \beta) \\
&= y_{\bullet\bullet}^2 / an + T_{yy} + (E_{xy})^2 / E_{xx} - y_{\bullet\bullet}^2 / an - (S_{xy})^2 / S_{xx} \\
&= S_{yy} - (S_{xy})^2 / S_{xx} - \left[E_{yy} - (E_{xy})^2 / E_{xx}\right] \quad \{\because T_{yy} = S_{yy} - E_{yy}\} \\
&= SS'_E - SS_E
\end{aligned}
$$

Note that above sum of square has (*an* – 2) – {*a*(*n*-1) -1} = *a* – 1 degrees of freedom.

It may be noted that SS_E is smaller than SS'_E [**because SS_E corresponds to full model that contains additional parameters $\{\tau_i\}$ and consequently explained variability will be more**] and that the quantity $SS'_E - SS_E$ will be small under the null hypothesis of no treatment effect, i.e. $\tau_i = 0$, else the same will be large. Therefore, the difference between SS'_E and SS_E, that is, $SS'_E - SS_E$, provides a sum of squares with *a*-1 degrees of freedom for testing the hypothesis of no treatment effects.

Consequently, to test $H_0 : \tau_i = 0$, compute

$$F_0 = \frac{(SS_E' - SS_E)/(a-1)}{SS_E/[a(n-1)-1]}$$

which, if the null hypothesis is true, is distributed as $F_{a-1,\ a(n-1)-1}$. Thus, we reject the null hypothesis of no treatment effect if $F_0 > F_{\alpha,\ a-1,a(n-1)-1}$. The P value approach could also be used.

Table 2: Analysis of Covariance as an "Adjusted" Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F₀
Regression	$(S_{xy})^2/S_{xx}$	1		
Treatments	$SS_E' - SS_E$	$a-1$	$MS_T = \dfrac{SS_E' - SS_E}{a-1}$	$\dfrac{MS_T}{MS_E}$
Error	SS_E	$a(n-1)-1$	$MS_E = \dfrac{SS_E}{a(n-1)-1}$	
Total	S_{yy}	$an-1$		

Note: $SS_E = E_{yy} - (E_{xy})^2 / E_{xx}$, $\quad SS_E' = S_{yy} - (S_{xy})^2 / S_{xx}$

It is instructive to examine the display in Table 2 above. In this table we have presented the analysis of covariance as an "adjusted" analysis of variance. In the source of variation column, the total variability is measured by S_{yy} with *an* - 1 degrees of freedom. The source of variation "regression" has the sum of squares is $(S_{xy})^2/S_{xx}$ with one degree of freedom. If there were no concomitant variable, we would have $S_{xy} = S_{xx} = E_{xy} = E_{xx} = 0$. Then the sum of squares for error would be simply E_{yy} [refer expression for *SS*_E] and the sum of squares for treatments would be $S_{yy} - E_{yy}$. [from the expression $R[\tau|\mu,\beta]$]

However, because of the presence of the concomitant variable, we must "adjust" S_{yy} and E_{yy} for the regression of y on x as shown in Table 2 to get a better estimate error sum of squares. The adjusted error sum of squares has $a(n-1)-1$ degrees of freedom instead of $a(n-1)$ degrees of freedom because an additional parameter (the slope β) is fitted to the data.

The computations are usually displayed in an analysis of covariance table such as Table 3 given below.

Table 3: Analysis of Covariance (Single Factor Experiment with One Covariate)

Source of Variation	Degrees of Freedom	Sum of Squares and Products			Adjusted for Regression of y on x		
		x	xy	y	Sum of Squares	Degrees of Freedom	Mean Square
Treatment	a - 1	T_{xx}	T_{xy}	T_{yy}			
Error	$a(n-1)$	E_{xx}	E_{xy}	E_{yy}	SS_E	$a(n-1)-1$	$MS_E = \dfrac{SS_E}{a(n-1)-1}$
Total	an - 1	S_{xx}	S_{xy}	S_{yy}	SS'_E	$an-2$	
Adjusted treatments					$SS'_E - SS_E$	$a-1$	$\dfrac{SS'_E - SS_E}{a-1}$

Note: $SS_E = E_{yy} - \left(E_{xy}\right)^2 / E_{xx}$, $\quad SS'_E = S_{yy} - \left(S_{xy}\right)^2 / S_{xx}$

Above layout conveniently summarizes all the required sums of squares and cross-products as well as the sums of squares for testing hypotheses about treatment effects. In addition to testing the hypothesis that there are no differences in the treatment effects, we frequently find it useful in interpreting the data to present the adjusted treatment means.

These adjusted means are computed as

$$\text{Adjusted } \bar{y}_{i\cdot} = \bar{y}_{i\cdot} - \hat{\beta}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}) \qquad i = 1, 2, \cdots, a$$

where $\hat{\beta} = E_{xy}/E_{xx}$. These adjusted treatment means are the least squares estimators of $\mu + \tau_i, \;\; i = 1,2, \cdots, a$ in the model (Equation 1).

The standard error of any adjusted treatment mean is

$$SE_{\text{adj } \bar{y}_{i\cdot}} = \left[MS_E\left(\frac{1}{n} + \frac{(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2}{E_{xx}}\right)\right]^{1/2} \quad \left[\text{Since } Cov\left(\bar{y}_{i\cdot}, \hat{\beta}\right) = 0\right]$$

Note: **Covariance of two linear combinations**:

Let $a = \sum_{i=1}^{n} a_i y_i$ and $c = \sum_{i=1}^{n} c_i y_i$. Moreover, suppose $V(y_i) = \sigma^2$ and y's are pairwise uncorrelated, i.e. $cov\left(y_i, y_j\right) = 0, i \neq j$. Then

$$cov(a, c) = \sigma^2 \sum a_i c_i$$

$$a = \bar{y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \qquad c = \hat{\beta} = \frac{E_{xy}}{E_{xx}} = \frac{1}{E_{xx}}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\left(x_{ij} - \bar{x}_{i\cdot}\right)$$

Therefore, $cov\left(\bar{y}_{i\cdot}, \hat{\beta}\right) = \sigma^2 \frac{1}{nE_{xx}} \sum_{i=1}^{a}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_{i\cdot}\right) = 0.$

Finally, we recall that the regression coefficient β in the full model (Equation 1) has been assumed to be nonzero. We may test the significance of regression, i.e. the hypothesis $H_0 : \beta = 0$ by using the following test statistic :

$$F_0 = \frac{\left(E_{xy}\right)^2 / E_{xx}}{MS_E}$$

Since, $SS_E = E_{yy} - \frac{\left(E_{xy}\right)^2}{E_{xx}}$, so, for full model, we get, $SS(\text{Reg}) = \frac{\left(E_{xy}\right)^2}{E_{xx}}$.

which under the null hypothesis is distributed as $F_{1,a(n-1)-1}$. Thus, we reject $H_0 : \beta = 0$, if $F_0 > F_{\alpha,1,a(n-1)-1}$.

EXAMPLE

Consider the experiment described at the beginning in Table 1. Three different machines produce a monofilament fibre for a textile company. The process engineer is interested in determining if there is a difference in the breaking strength of the fibre produced by the three machines. However, the strength of a fibre is related to its diameter, with thicker fibres being generally stronger than thinner ones. A random sample of five fibre specimens is selected from each machine.

The scatter diagram of breaking strength versus the fibre diameter (Figure 1) shows a strong suggestion of a linear relationship between breaking strength and diameter, and it seems appropriate to remove the effect of diameter on strength by an analysis of covariance. Assuming that a linear relationship between breaking strength and diameter is appropriate, the model is

$$y_{ij} = \mu + \tau_i + \beta\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) + \varepsilon_{ij} \qquad \begin{cases} i = 1, 2, 3 \\ j = 1, 2, \ldots, 5 \end{cases}$$

Let us first make following table of machine-wise totals and averages.

Table 4: Summary of Breaking Strength Data (y = strength in pounds and x = diameter in 10^{-3} inches)

	Machine 1		Machine 2		Machine 3	
	y	x	y	x	y	x
	36	20	40	22	35	21
	41	25	48	28	37	23
	39	24	39	22	42	26
	42	25	45	30	34	21
	49	32	44	28	32	15
TOTAL →	**207**	**126**	**216**	**130**	**180**	**106**
AVERAGE →	**41.04**	**25.2**	**43.2**	**26**	**36**	**21.2**

Grand Total for y ($y_{..}$)	603
Grand Total for x ($x_{..}$)	362

Using the formulae stated earlier, we compute the following.

$$S_{yy} = 346.40 \quad S_{xx} = 261.73 \quad S_{xy} = 282.60$$
$$T_{yy} = 140.40 \quad T_{xx} = 66.13 \quad T_{xy} = 96.00$$
$$E_{yy} = 206.00 \quad E_{xx} = 195.60 \quad E_{xy} = 186.60$$

$SS_E' = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = 41.27$ with $an - 2 = 15 - 2 = 13$ degrees of freedom.

$SS_E = E_{yy} - \frac{(E_{xy})^2}{E_{xx}} = 27.99$ with $a(n-1) - 1 = 11$ degrees of freedom.

So, the sum of squares for testing $H_0 : \tau_i = 0$ is

$$SS_E' - SS_E = 41.27 - 27.99 = 13.28$$

with a – 1 = 3 – 1 = 2 degrees of freedom.

The calculations are summarized in Table 5 below.

Table 5: Analysis of Covariance for Breaking Strength data

Source of Variation	Degrees of Freedom	Sum of Squares and Sum of Products			Adjusted for Regression of *y* on *x*		
		x	*xy*	*y*	Sum of Squares	Degrees of Freedom	Mean Square
Machines	2	66.13	96.00	140.40			
Error	12	195.60	186.60	206.00	27.99	11	2.54
Total	14	261.73	282.60	346.40	41.27	13	
Adjusted treatments					13.28	2	6.64

To test the hypothesis that machines differ in the breaking strength of fibre produced, that is, $H_0 : \tau_i = 0$, compute the test statistic

$$F_0 = \frac{(SS'_E - SS_E)/(a-1)}{SS_E/[a(n-1)-1]} = \frac{13.28/2}{27.99/11} = 2.61.$$

Comparing this to $F_{0.05,2,11} = 3.98$, we find that the null hypothesis cannot be rejected. So, there is no strong evidence that the fibre produced by three machines differ in breaking strength.

The estimate of the regression coefficient is computed as

$$\hat{\beta} = \frac{E_{xy}}{E_{xx}} = \frac{186.6}{195.6} = 0.9540\ .$$

We may now test the hypothesis $H_0 : \beta = 0$ using the statistic given below

$$F_0 = \frac{(E_{xy})^2/E_{xx}}{MS_E} = \frac{186.6^2/195.6}{2.54} = 70.08,$$

which is clearly much larger than $F_{0.01,1,11} = 9.65$ and we reject the hypothesis. Therefore, there is a linear relationship between breaking strength and diameter, and the adjustment provided by the analysis of covariance was necessary.

The adjusted treatment means may be computed from equation given earlier. These adjusted means are

Adjusted $\bar{y}_{1\bullet} = \bar{y}_{1\bullet} - \hat{\beta}(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})$= 41.40 – 0.9540(25.20 – 24.13) = 40.38,

Adjusted $\bar{y}_{2\bullet}$ = 43.20 - 0.9540(26.00-24.13) = 41.42, and

Adjusted $\bar{y}_{3\bullet}$ = 36.00 – 0.9540(21.20 – 24.13) = 38.80.

Comparing the adjusted treatment means with the unadjusted treatment means (i.e. $\bar{y}_{i\cdot}$), we note that the adjusted means are much closer together, another indication that the covariance analysis was necessary.

A basic assumption in the analysis of covariance is that the treatments do not influence the covariate *x* and the technique removes the portion of variation in the response that could be due to the variations in the *x*. However, if the variability in $\bar{x}_{i\cdot}$ is partly due to the treatments, then analysis of covariance removes part of the treatment effect too. Thus, we must be reasonably sure that the treatments do not affect the values x_{ij}. In some experiments this may be obvious from the nature of the covariate, whereas in others it may not be so. In our example, there may be a difference in fibre diameter $\left(x_{ij}\right)$ between the three machines. In such cases, Cochran and Cox in their book on Experimental Design, suggest that an analysis of variance on the x_{ij} values may be helpful in determining the validity of this assumption. For our problem, this procedure yields

$$F_0 = \frac{T_{xx}/(a-1)}{E_{xx}/\{a(n-1)\}} = \frac{66.13/2}{195.6/12} = \frac{33.07}{16.30} = 2.03$$

which is less than $F_{0.05,2,12} = 3.89$, so there is no reason to believe that machines produce fibres with different average diameter.

Exercise:

Using General Regression Significance Test, develop ANOVA testing procedure for a Single-Factor fixed effects model.

[**Hint**:

$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad i = 1(1)a$ and $j = 1(1)n$

Determine $R(\mu, \tau)$ for full model and $R(\mu)$ for the model restricted by null hypothesis.

SS Treatment will be $R(\tau|\mu) = R(\mu, \tau) - R(\mu)$ and SS Error = $\sum\sum y_{ij}^2 - R(\mu, \tau)$, both with appropriate degrees of freedom.]

DIAGNOSTIC CHECKING – RESIDUAL ANALYSIS

For the covariance model, the residuals are

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

where the fitted values are

$$\begin{aligned}\hat{y}_{ij} &= \hat{\mu} + \hat{\tau}_i + \hat{\beta}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) \\ &= \bar{y}_{\bullet\bullet} + \left[\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet} - \hat{\beta}\left(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet}\right)\right] + \hat{\beta}\left(x_{ij} - \bar{x}_{\bullet\bullet}\right) \\ &= \bar{y}_{i\bullet} + \hat{\beta}\left(x_{ij} - \bar{x}_{i\bullet}\right)\end{aligned}$$

Table 6: Observed values, fitted values and residuals.

Observed Value, y_{ij}	Fitted Value, $\hat{y}_{ij}$	Residual $e_{ij} = y_{ij} - \hat{y}_{ij}$
36	36.4392	-0.4392
41	41.2092	-0.2092
39	40.2552	-1.2552
42	41.2092	0.7908
49	47.8871	1.1129
40	39.3840	0.6160
48	45.1079	2.8921
39	39.3840	-0.3840
45	47.0159	-2.0159
44	45.1079	-1.1079
35	35.8092	-0.8092
37	37.7171	-0.7171
42	40.5791	1.4209
34	35.8092	-1.8092
32	30.0852	1.9148

The residuals are plotted versus the fitted values in Figure 2, versus the covariate *x* in Figure 3 and versus the machines in Figure 4, whereas a normal probability plot of the residuals is shown in Figure 5. These plots do not reveal any major departures from the assumptions, so we conclude that the covariance model (Equation 1) is appropriate for the breaking strength data.

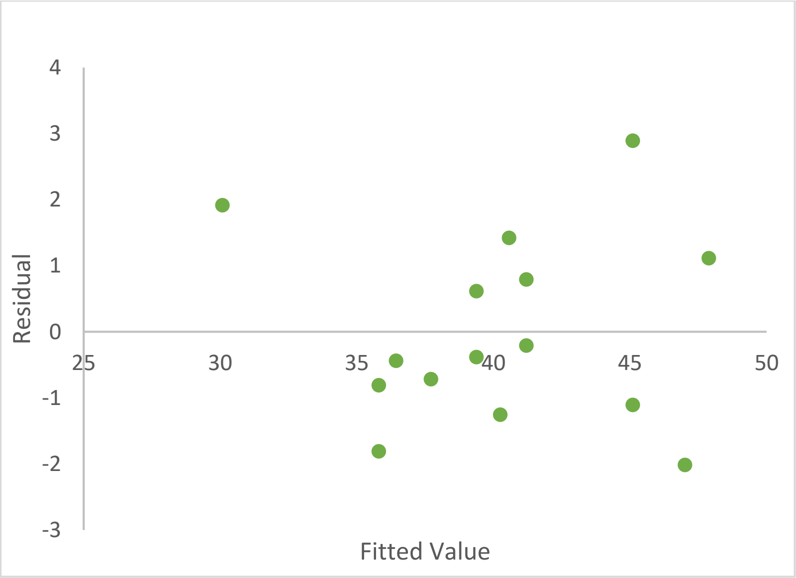


Figure 2: Plot of residuals versus fitted values

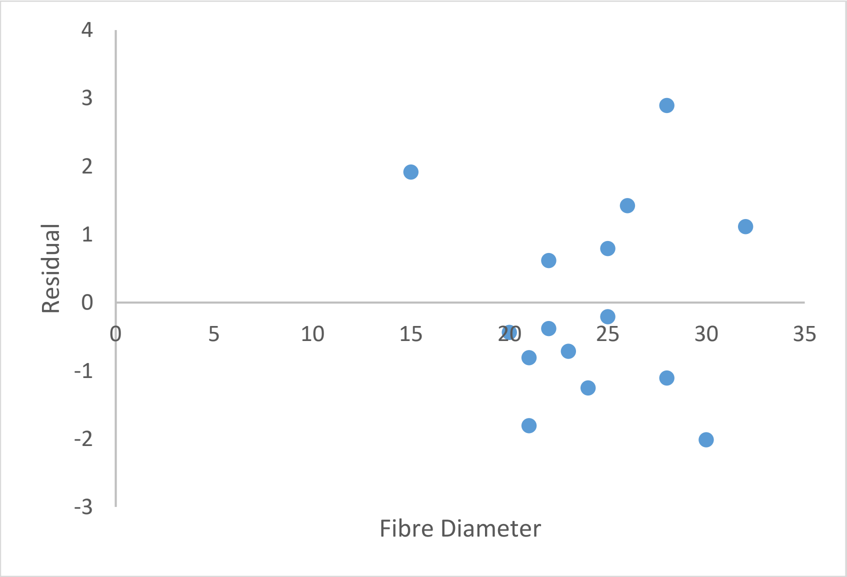


Figure 3: Plot of residuals versus fibre diameter, *x*

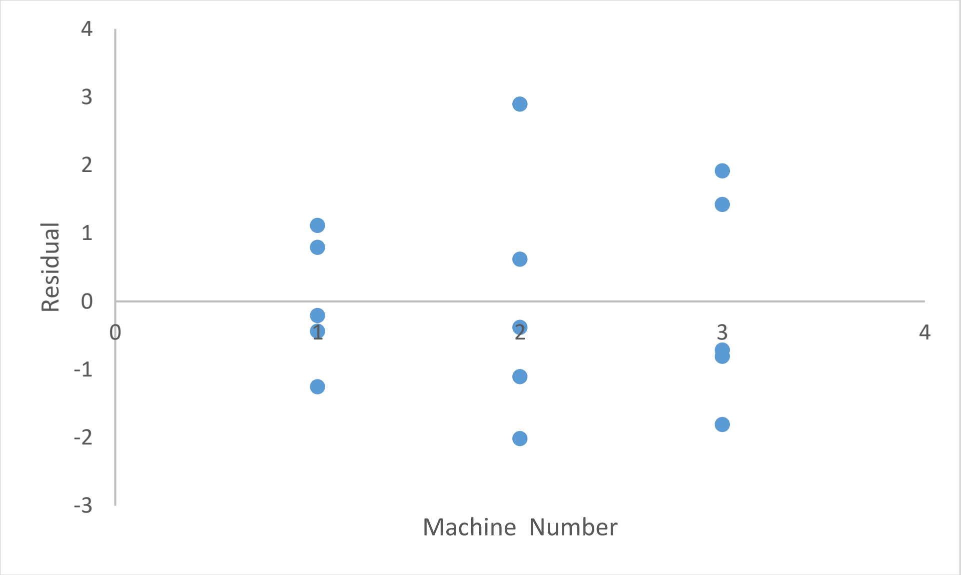


Figure 4: Plot of residuals versus machine number

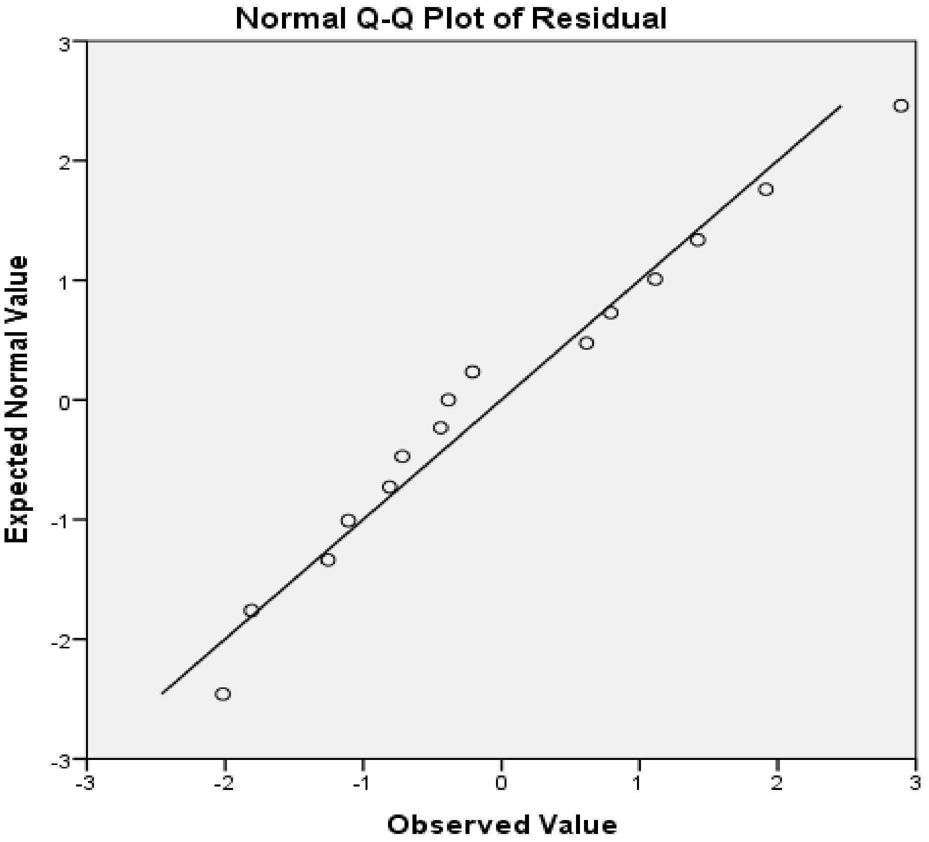


Figure 5: Normal Probability Plot of residuals

ANOVA not ANCOVA

It is interesting to note what would have happened in this experiment if an analysis of covariance had not been performed, that is, if the breaking strength data (y) had been analysed as a single-factor experiment in which the covariate x was ignored. The analysis of variance of the breaking strength data is shown in Table 7.

Table 7: Incorrect analysis of Breaking Strength data as a Single-Factor experiment

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P – value
Machines	140.40	2	70.20	4.09	0.0442
Error	206.00	12	17.17		
Total	**346.00**	**14**			

We would conclude, based on this analysis that machines differ significantly in the strength of fibre produced. This is exactly the opposite conclusion reached by the covariance analysis.

If we suspect that the machines differ significantly in their effect on fibre strength, then we would try to control the strength output of the three machines. However, in this problem the machines do not differ in the strength of fibre produced after the linear effect of fibre diameter in removed.

It would be helpful to reduce the within-machine fibre diameter variability since this would probably reduce the strength variability in the fibre.