8. Regression and Analysis of Variance

The analysis of variance (ANOVA) is a statistical method developed by
The analysis of variance (ANOVA) is a statistical method developed by
R A Fisher for the analysis of experimental data. Initially it was applied to the
analysis of agricultural experiments (use of various fertilisers, various seeds),
but soon its application expanded to many other fields of scientific research.
With the method of analysis of variance we can break down the total variance
of a variable into additive components which may be attributed to various,
separate factors. These factors are the 'causes' or 'sources' of variation of the
variable being analysed. The method, when applied to experimental data,
assumes a certain design of the experiment, which determines the number of
the relevant factors (or causes) of variation and the logical significance of each
one of them. For example, assume that we have twenty plots of land on which
we cultivate wheat, and we want to study the yield per unit of land. We use
different seeds, different fertilisers and different systems of irrigation. Thus the
variation in yields may logically be attributed to the three factors:

X_1 = type of seed
X_2 = type of fertiliser
X_3 = type of irrigation

With the method of analysis of variance we may break down the total variation
in yield into three separate components: a component due to X_1, another due
to X_2 and a third due to X_3.

From this definition of the analysis of variance it should be clear that this
method is conceptually the same as regression analysis. In regression analysis
also the aim is to determine the factors which cause the variation of the
dependent variable. We saw that the total variation in the dependent variable
is split into two components: the variation explained by the regression line (or
regression plane), and the unexplained variation, shown by the scatter of
points around the regression line. Furthermore, the multiple correlation
coefficient was seen to represent the proportion of total variation explained by
the regression line (or regression plane). R^2 was found to be equal to additive
components, each corresponding to a relevant explanatory variable. However,
there are significant differences between the two methods. The main difference
is that regression analysis provides numerical values for the influence of the
various explanatory factors on the dependent variable, in addition to the
information concerning the breaking down of the total variance of Y into
additive components, while the analysis of variance provides only the latter
type of information.

Both the analysis of variance and regression analysis have as their objective
the determination of the various factors which cause variations of the dependent
variable. This resemblance has led to the combination of the two methods in
most scientific fields. In particular, the method of analysis of variance is used
in regression analysis for conducting various tests of significance, the most
important being:

(1) The test of the overall significance of the regression.

(2) The test of the significance of the improvement in fit obtained by the
introduction of additional explanatory variables in the function. This test is
formally equivalent to the t test developed in Chapter 5.

(3) The test of the equality of coefficients obtained from different samples.

(4) The test of the extra-sample performance of a regression, or test of the
stability of the regression coefficients.

(5) The test of restrictions imposed on coefficients of a function.

In this chapter we shall examine the use of analysis of variance ideas in
regression analysis for carrying out the above tests. In order to understand them
it is necessary to begin with a short description of the method of analysis of
variance, as a statistical method in its own right.

8.1. THE METHOD OF ANALYSIS OF VARIANCE AS A STATISTICAL METHOD

The aim of this method is to split the total variation of a variable (around
its mean) into components which may be attributed to specific (additive) causes.
To simplify the analysis we will assume that there is only one systematic factor
which influences the variable being studied. Any variation not accounted for by
this (explanatory) factor is assumed to be random (or chance) variation, due to
various random happenings. We have a series of values of a variable Y and the
corresponding values of the (explanatory) variable X. The analysis of variance
method concentrates on the values of Y and studies their variation. The values
of X are used only for dividing the values of Y into sub-groups, sub-samples;
for example one group (or sample) corresponding to large values of X and one
group (or sample) corresponding to small values of X.

For each sub-sample we estimate the mean-value of Y, obtaining a set of
means. If X (which is the basis of the classification of the Y's into the sub-
samples) is an important cause of variation in Y (an important explanatory
variable) the difference between the means of the sub-samples will be large:
this would be shown by a large dispersion of the means of sub-samples $\bar{Y}_i$'s
around the common mean $\bar{Y}$, that is, by a large variance of the distribution
of the means. On the contrary, if X is not an important source of variation of
Y, the difference between the means of the sub-samples will be small, a fact
that would be reflected in a small variance of the distribution of sampling
means ($\bar{Y}_i$) around the common mean $\bar{Y}$:

(a) The importance of X as a cause of variation (in Y) is judged from the
difference between the means of sub-samples ($\bar{Y}_i$'s), formed on the basis of
the values of X.

(b) The difference between the means is reflected in the value of the variance of the distribution of the sample means.

Hence the difference between the means may be studied and tested with two estimates of the population variance of Y. One estimate of σ_Y^2 is obtained by pooling the variances of the sub-samples, and the other is obtained from the expression of the sampling distribution (the distribution of $\bar{Y}$). Whatever the relationship of the data being studied, the method of analysis of variance reduces to the estimation of two variances, and the comparison of these variances in order to establish whether the difference between them is statistically significant, or whether it is due to chance, in which case we conclude that there is no real difference between the variance-estimates.

The comparison of any two variances is implemented by the F statistic and the F tables (reproduced on pp. 663–4). The F statistic is the ratio of any two *independent* estimates of variances, which have been obtained from sample data.[1] Each estimate involves some loss of degrees of freedom. If we have any two independent variance estimates obtained with ν_1 and ν_2 degrees of freedom respectively, their ratio has the F distribution with ν_1 and ν_2 degrees of freedom. For this reason F is called the *variance ratio*. (See Appendix I.) The letter F stands for the name of Fisher who invented this statistic.

If the two variance estimates are close to each other their ratio will approach the value of one. The greater the discrepancy (difference) between the two variances the greater is the value of the F ratio. Thus, in general, high values of F suggest that the difference between the two variances is significant, or the rejection of the null hypothesis, which assumes no significant difference between the two variances.

We will illustrate the method of analysis of variance with an example.

Test of the difference between means

Suppose three different types of petrol are used for running a car: type A rated at 90 octane, type B rated at 95 octane and type C at 100 octane. We wish to test whether these different types of petrol give the same consumption per mile, that is, we want to compare the consumption performance of the three brands of petrol. Suppose that we use each brand for ten days and we measure the miles per gallon of petrol. Thus we obtain three samples of size 10 for each brand. The observations, shown in table 8.1, report miles per gallon of petrol.

The above data may be interpreted as three random samples of size $n_1 = n_2 = n_3 = 10$, with means $\bar{Y}_1 = 33$, $\bar{Y}_2 = 38$ and $\bar{Y}_3 = 46$ miles per gallon of petrol. Our problem is to establish whether the difference between these means is significant or whether it may be attributed to chance.

We shall assume that the samples are drawn from three populations which have a normal distribution (or approximately normal) with means μ_1, μ_2 and μ_3 respectively and with equal standard deviation σ. This assumption implies

[1] See Appendix I.

Table 8.1

Sample 1 *Brand A* $n_1 = 10$	*Sample 2* *Brand B* $n_2 = 10$	*Sample 3* *Brand C* $n_3 = 10$	*Total observations* $N = n_1 + n_2 + n_3$
32	35	44	32
30	38	46	30
35	37	47	35
33	40	47	33
35	41	46	35
34	35	43	34
29	37	47	29
32	41	45	32
36	36	48	36
34	40	47	34
			35
			38
			37
			40
			41
			35
			37
			41
			36
			40
			44
			46
			47
			47
			46
			43
			47
			45
			48
			47
$\Sigma Y_{1i} = 330$	$\Sigma Y_{2i} = 380$	$\Sigma Y_{3i} = 460$	$\sum_j \sum_i Y_{ji} = 1170$
$\bar{Y}_1 = \frac{\Sigma Y_{1i}}{n_1} = 33$	$\bar{Y}_2 = \frac{\Sigma Y_{2i}}{n_2} = 38$	$\bar{Y}_3 = \frac{\Sigma Y_{3i}}{n_3} = 46$	$\bar{Y} = \frac{\Sigma \Sigma Y_{ji}}{N} = 39$
$S_1^2 = \frac{\sum_i^{n_1} (Y_{1i} - \bar{Y}_1)^2}{n_1} = \frac{46}{10}$, $S_1^2 = 4{\cdot}6$	$S_2^2 = \frac{\sum_i^{n_2} (Y_{2i} - \bar{Y}_2)^2}{n_2} = \frac{50}{10}$, $S_2^2 = 5{\cdot}0$	$S_3^2 = \frac{\sum_i^{n_3} (Y_{3i} - \bar{Y}_3)^2}{n_3} = \frac{22}{10}$, $S_3^2 = 2{\cdot}2$	

that although the different octane content of the three brands of petrol may affect the average consumption of petrol, it would not affect the dispersion (variance) of the mileages around the means. In other words, if we take a large number of observations for each brand of petrol, the three distributions which we would get would be close to normal curves having the same standard deviation σ. We want to know whether there is any significant difference between the means of the populations, μ_1, μ_2 and μ_3. We want to test the null hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3$$

against the alternative hypothesis

$$H_1: \mu_j \text{ not all equal.}$$

If the three means are the same, that is if the null hypothesis is true, the three populations may be considered as one large population with mean $\mu (= \mu_1 = \mu_2 = \mu_3)$ and standard deviation σ, that is,

$$Y \sim N(\mu, \sigma)$$

and the three samples may be considered as samples drawn from this one large population.

Applying the basic sampling theorems[1] we may write the following distributions for the sample means $\overline{Y}_1$, $\overline{Y}_2$, $\overline{Y}_3$

$$\overline{Y}_1 \sim N(\mu, \sigma^2_{\overline{Y}_1}) \sim N\left(\mu, \frac{\sigma^2}{n_1}\right)$$

$$\overline{Y}_2 \sim N(\mu, \sigma^2_{\overline{Y}_2}) \sim N\left(\mu, \frac{\sigma^2}{n_2}\right)$$

$$\overline{Y}_3 \sim N(\mu, \sigma^2_{\overline{Y}_3}) \sim N\left(\mu, \frac{\sigma^2}{n_3}\right)$$

We said that under the null hypothesis ($\mu_1 = \mu_2 = \mu_3$) we may consider the three populations as forming a large population

$$Y \sim N(\mu, \sigma^2)$$

An estimate of the common mean μ may be computed from the enlarged sample $n_1 + n_2 + n_3 = N = 30$. From the data of table 8.1 we obtain

$$\hat{\mu} = \frac{\Sigma Y_i}{N} = \frac{\sum_j^k \sum_i^{n_k} Y_{ji}}{N} = \frac{1170}{30} = 39 = \overline{Y}$$

[1] See Appendix I. If a variable X is normally distributed, that is

$$X \sim N(\mu, \sigma^2)$$

then the sample means in repeating sampling will also have a normal distribution

$$\overline{X}_i \sim N(\mu, \sigma^2_{\overline{X}}) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

An estimate of the population variance σ^2 may be obtained in two ways.

Firstly. An unbiased estimator of the population variance may be obtained from the expression

$$\hat{\sigma}^2 = \frac{\sum_j^k n_j(\overline{Y}_j - \overline{Y})^2}{k-1} \tag{8.1}$$

where k is the number of samples.

Proof. This expression is derived from the relationship between the population variance σ^2 and the variance of the sampling distribution:

$$\sigma^2_{\overline{Y}_j} = \frac{\sigma^2}{n_j} \quad \text{or} \quad \sigma^2 = \sigma^2_{\overline{Y}_j} \cdot n_j$$

In our example we have three samples and from each one of them we may obtain a separate estimate of σ^2:

$$\hat{\sigma}^2_1 = n_1 \cdot \sigma^2_{\overline{Y}_1} = n_1(\overline{Y}_1 - \overline{Y})^2$$
$$\hat{\sigma}^2_2 = n_2 \cdot \sigma^2_{\overline{Y}_2} = n_2(\overline{Y}_2 - \overline{Y})^2$$
$$\hat{\sigma}^2_3 = n_3 \cdot \sigma^2_{\overline{Y}_3} = n_3(\overline{Y}_3 - \overline{Y})^2$$

where $\overline{Y}$ is the common (pooled) mean.

Taking the weighted average of these estimates we obtain

$$\hat{\sigma}^2 = \frac{1}{3} \sum_j^k n_j(\overline{Y}_j - \overline{Y})^2$$

For an unbiased estimate we use the degrees of freedom $3 - 1 = 2$, or in general $k - 1$, if we have k samples. Thus the first estimate of the population variance becomes

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^k n_j(\overline{Y}_j - \overline{Y})^2}{k-1}$$

It should be clear that this estimate of the population variance is obtained from the differences between the sample means ($\overline{Y}_j$) and the common population mean ($\overline{Y}$). Recall that the sample means are unbiased estimates of the means μ_1, μ_2, μ_3. The null hypothesis was $\mu_1 = \mu_2 = \mu_3 = \mu$. Hence if this hypothesis should be true, the sample means $\overline{Y}_1$, $\overline{Y}_2$, $\overline{Y}_3$ should not differ significantly from each other and also from the overall mean $\overline{Y}$. This implies that if the null hypothesis is not true, we should expect that the sample means, $\overline{Y}_1$, $\overline{Y}_2$, $\overline{Y}_3$ should also differ considerably from each other and from the common (pooled) mean $\overline{Y}$: the difference between these means would be larger than what may be attributed to chance. This in turn implies that the estimate $\hat{\sigma}^2$ of the population variance will be large if the null hypothesis is not true, because $\hat{\sigma}^2$ was computed from the differences $(\overline{Y}_j - \overline{Y})^2$. Thus the estimate $\hat{\sigma}^2$ is the crucial element of the test of difference between means of various samples. From the way it is estimated it reflects the variation between the sample means and it is called *'variation between'*.

To conduct our test it suffices to compare this estimate with the true population variance σ^2, and reject the null hypothesis if the divergence between $\hat{\sigma}^2$ and σ^2 is large. However, in our example, as in most actual problems of the real world, the true σ^2 is unknown, and we have to obtain another independent estimate from the sample data.

Secondly. An estimate of the population variance σ^2 may be obtained by pooling together the various sample variances. The appropriate formula is

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + \ldots + n_k s_k^2}{(n_1 + n_2 + \ldots + n_k) - k}, \tag{8.2}$$

where s_j are the sample variances and n_j the sample sizes. (See Yamane, *Statistics*, p. 504.) Note that

$$\left.\begin{aligned} n_1 s_1^2 &= n_1 \frac{\Sigma(Y_{1i} - \overline{Y}_1)^2}{n_1} = \Sigma(Y_{1i} - \overline{Y}_1)^2 \\ &\vdots \\ n_k s_k^2 &= n_k \frac{\sum_i^{n_k}(Y_{ki} - \overline{Y}_k)^2}{n_k} = \Sigma(Y_{ki} - \overline{Y}_k)^2 \end{aligned}\right\} \tag{8.3}$$

Thus $n_1 s_1^2 + n_2 s_2^2 + \ldots + n_k s_k^2$ gives the total sum of squared deviations of all k samples, and the 'pooled-variance' expression can be considered as an operation of combining all samples into one large sample and estimating the population variance. Substituting 8.3 in 8.2 we obtain

$$\hat{\sigma}^2 = \frac{\sum_i^{n_1}(Y_{1i} - \overline{Y}_1) + \sum_i^{n_2}(Y_{2i} - \overline{Y}_2)^2 + \ldots + \sum_i^{n_k}(Y_{ki} - \overline{Y}_k)^2}{N-k},$$

where $N = n_1 + n_2 + \ldots + n_k$. Using the double summation notation (see Appendix I, p. 519), we have

$$\hat{\sigma}^2 = \frac{\sum_j^k \sum_i^{n_k}(Y_{ji} - \overline{Y}_j)^2}{N-k}. \tag{8.4}$$

This estimate of the population variance is obtained from the sample variances which reflect the variation *within* each sample. The sample variances do not depend on the null hypothesis, they are not affected by differences between the sample means ($\overline{Y}_1$, $\overline{Y}_2$, $\overline{Y}_3$). In other words even if the means are significantly different, in which case we will have three populations each having its own different mean (μ_1, μ_2, μ_3), all these populations would have (by assumption) the same variance σ^2, and hence $\hat{\sigma}^2$ would be an unbiased estimate of the variance σ^2 of the 'pooled' population.

$\hat{\sigma}^2$ is based on the variation *within* the sample values (Y_i's of each sample), and is called '*within variation*'. Now note that the variation of the values of Y_i in each sample are chance variations, so that the estimate $\hat{\sigma}^2$ may be considered

as a measure of the variation in the values of Y_i's which may be attributed to chance.

We now have two unbiased estimates of the population variance σ^2:

Estimate (1) reflects the *variation between the sample means*, and depends on the validity of the null hypothesis.

Estimate (2) reflects the *variation of* Y_i*'s within the samples*, and is independent of the null hypothesis.

It can be shown[1] that the two estimates are independent, so that their ratio has an F distribution with $\nu_1 = k - 1$ and $\nu_2 = N - k$ degrees of freedom:

$$F^* = \frac{\left[\sum_{j=1}^{k} n_j(\overline{Y}_j - \overline{Y})^2\right] \Big/ (k-1)}{\left[\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ji} - \overline{Y}_j)^2\right] \Big/ (N-k)}$$

where: n_j = size of the jth sample

$N = \sum_{j=1}^{k} n_j$ = size of the 'pooled' (enlarged) sample

k = number of samples.

The variance ratio may be shown schematically as

$$F^* = \frac{\text{estimated variance from 'between'-the-means variation}}{\text{estimated variance from 'within'-the-samples variation}}$$

When the means (μ_1, μ_2, μ_3) are not equal the estimated variance from the 'between'-the-means differences will be large and hence the variance ratio F^* will become large. If the null hypothesis is true the observed variance ratio will approximate the value of one: the observed difference in the means $\overline{Y}_1$, $\overline{Y}_2$, $\overline{Y}_3$ in this case is not significant and may well be attributed to chance; thus the estimate appearing in the numerator of F^* will be really estimating the same unknown population variance as the denominator is also estimating.

The observed F^* variance ratio is compared with the theoretical value of F (with a chosen level of significance, e.g. the 5 per cent level), which is found from the F-table (pp. 663–4) with $\nu_1 = (k - 1)$ and $\nu_2 = (N - k)$ degrees of freedom. The theoretical (or critical) value of F is the value of F that defines the critical region of the test at the chosen level of significance.

If $F^* > F$ we reject the null hypothesis, i.e. we accept that the difference between the means is significant. From this evidence we may infer that the populations, from which the samples are drawn, do differ.

If $F^* < F$ we accept the null hypothesis, i.e. we accept that the sample means are not significantly different. In this event we may say that the sample data provide evidence that there is no significant difference between the means of the populations from which the samples are drawn.

[1] See G. Yule and M. Kendall, *An Introduction to the Theory of Statistics*, 14th edition, New York, Hefner, 1950, p. 507.

In our example we have the following results:

(1) The 'between' variance estimate is

$$\hat{\sigma}^2 = \frac{\sum_j^k n_j(\overline{Y}_j - \overline{Y})^2}{k-1}.$$

$$= \frac{n_1(\overline{Y}_1 - \overline{Y})^2 + n_2(\overline{Y}_2 - \overline{Y})^2 + n_3(\overline{Y}_3 - \overline{Y})^2}{3-1}$$

$$= \frac{10(33-39)^2 + 10(38-39)^2 + 10(46-39)^2}{2} = 430$$

(2) The 'within' variance estimate is

$$\hat{\hat{\sigma}}^2 = \frac{\sum_j^k \sum_i^{n_j} (Y_{ji} - \overline{Y}_j)^2}{N-k}$$

$$= \frac{\sum_1^{10} (Y_{1j} - \overline{Y}_1)^2 + \sum_1^{10} (Y_{2j} - \overline{Y}_2)^2 + \sum_1^{10} (Y_{3j} - \overline{Y}_3)^2}{30-3}$$

$$= \frac{46+50+22}{27} = \frac{118}{27} \approx 4{\cdot}37$$

(3) The observed variance ratio is

$$F^* = \frac{\hat{\sigma}^2}{\hat{\hat{\sigma}}^2} = \frac{430}{4{\cdot}37} = 98{\cdot}39 \approx 98{\cdot}4.$$

(4) The theoretical value of F at the 5 per cent level of significance with $\nu_1 = k - 1 = 2$ and $\nu_2 = N - k = 27$ degrees of freedom is found from the F-tables (pp. 663–4)

$$F_{0.05} = 3{\cdot}37$$

(5) Since $F^* > F_{0.05}$ we reject the null hypothesis, that is, we accept that there is a significant difference in the average mileage obtained from the three types of petrol.

The above test may be examined in another way, which will systematise the analysis of variance method. We may obtain a third estimate of the population variance, σ^2, by using the enlarged sample, formed from the three sub-samples. The unbiased estimate will be

$$*\sigma^2 = \frac{\sum_i^N (Y_i - \overline{Y})^2}{N-1} = \frac{\sum_{j=1}^k \sum_{n=1}^{n_j} (Y_{ji} - \overline{Y})^2}{N-1},$$

where $N - 1$ = degrees of freedom for the estimate $*\sigma^2$. If we take the numerators

of the three estimates of the population variance ($*\sigma^2, \hat{\sigma}^2, \hat{\hat{\sigma}}^2$), we may establish the following relationship between these terms:

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ji} - \overline{Y})^2 = \sum_{j=1}^k n_j(\overline{Y}_j - \overline{Y})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ji} - \overline{Y}_j)^2$$

that is

$$\begin{bmatrix}\text{Total sum of}\\ \text{squared deviations}\end{bmatrix} = \begin{bmatrix}\text{Sum of squares}\\ \text{between groups}\end{bmatrix} + \begin{bmatrix}\text{Sum of squares}\\ \text{within groups}\end{bmatrix}$$

or

$$\begin{bmatrix}\text{Total variation}\\ \text{in } Y\end{bmatrix} = \begin{bmatrix}\text{Between}\\ \text{variation}\end{bmatrix} + \begin{bmatrix}\text{Within}\\ \text{variation}\end{bmatrix}$$

Proof: We start from the term on the left-hand side and we form the identity

$$(Y_{ji} - \overline{Y}) = (Y_{ji} - \overline{Y}) + \overline{Y}_j - \overline{Y}_j$$

or

$$(Y_{ji} - \overline{Y}) = (Y_{ji} - \overline{Y}_j) + (\overline{Y}_j - \overline{Y})$$

Squaring both sides, we have

$$(Y_{ji} - \overline{Y})^2 = (Y_{ji} - \overline{Y}_j)^2 + (\overline{Y}_j - \overline{Y})^2 + 2(Y_{ji} - \overline{Y}_j)(\overline{Y}_j - \overline{Y})$$

Summing over all values, we find

$$\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ji} - \overline{Y})^2 = \sum_j \sum_i (Y_{ji} - \overline{Y}_j)^2 + \sum_j \sum_i (\overline{Y}_j - \overline{Y})^2 + 2 \sum_j \sum_i (Y_{ji} - \overline{Y}_j)(\overline{Y}_j - \overline{Y})$$

The last term of this expression is equal to zero, since

$$2 \sum_j \sum_i (Y_{ji} - \overline{Y}_j)(\overline{Y}_j - \overline{Y}) = 2 \sum_j [(\overline{Y}_j - \overline{Y}) \sum_i (Y_{ji} - \overline{Y}_j)]$$

and given that $\sum_i (Y_{ji} - \overline{Y}_j) = 0$, because it is the sum of the deviations within each group (sample).

Therefore

$$\sum_j^k \sum_i^{n_j} (Y_{ji} - \overline{Y})^2 = \sum_j^k \sum_i^{n_j} (Y_{ji} - \overline{Y}_j)^2 + \sum_j^k n_j(\overline{Y}_j - \overline{Y})^2$$

Total = Within + Between

This expression shows how the total sum of squared deviations in Y (in all the groups taken together) is partitioned into two parts: one part of the total variation of Y is due to the difference between the means (octane ratings of the three types of petrol in our example) and the other part is due to chance (for example rain, mood of the driver of the car, etc.). Note that this partitioning of the total variation into additive components holds irrespective of whether the null hypothesis ($\mu_1 = \mu_2 = \mu_3$) holds or not. In our example the total variation in Y's around the common mean ($\overline{Y} = 39$) is

$$\sum_j^3 \sum_i^{10} (Y_{ji} - \overline{Y})^2 = 978$$

the between sum of squares is

$$\sum_{j}^{k} n_j(\bar{Y}_j - \bar{Y})^2 = 860$$

the within sum of squares is

$$\sum_{j}^{k}\sum_{i}^{n_j}(Y_{ji} - \bar{Y}_j)^2 = 118$$

Clearly

$$978 = 860 + 118$$
$$\text{Total} = \text{Between} + \text{Within}$$

that is, the partitioning of the total variance holds despite the refutation of the null hypothesis.

We may further establish a relationship between the degrees of freedom in each of the three estimates of the population variance.

(a) The degrees of freedom for the overall variance $\hat{\sigma}^2$ is $N - 1$.

(b) The degrees of freedom for the estimate based on the 'between' (the means) difference, $\hat{\sigma}^2$, is $k - 1$.

(c) The degrees of freedom for the estimate based on the 'within' (the groups) variation, $\hat{\sigma}^2$, is $N - k$.

It is easy to see that

$$\underset{\text{Total}}{(N-1)} = \underset{\text{Within}}{(N-k)} + \underset{\text{Between}}{(k-1)}$$

(For a proof of this result see Yamane, *Statistics*, p. 677.)

With the above information on the partitioning of the total sum of squares (total variation in Y) and the various degrees of freedom we may form the *Analysis of Variance Table* (table 8.2).

Table 8.2. Analysis of variance table

Source of variation (1)	*Sum of squares* (2)	*Degrees of freedom* (3)	*Mean square* (4) = (2) : (3)	*F* (5)
Between-the-means	$\sum_j^k n_j(\bar{Y}_j - \bar{Y})^2$	$\nu_1 = (k-1)$	$\dfrac{\sum_j n_j(\bar{Y}_j - \bar{Y})^2}{k-1}$	$F^* = \dfrac{\sum_j n_j(\bar{Y}_j - \bar{Y})^2/(k-1)}{\sum_j\sum_i (Y_{ji} - \bar{Y}_j)^2/(N-k)}$
Within-the-samples	$\sum_j^k\sum_i^{n_j}(Y_{ji} - \bar{Y}_j)^2$	$\nu_2 = (N-k)$	$\dfrac{\sum_j\sum_i (Y_{ji} - \bar{Y}_j)^2}{N-k}$	
Total variation	$\sum_j^k\sum_i^{n_j}(Y_{ji} - \bar{Y})^2$	$(N-1)$		**F from tables with** $\nu_1 = k-1$, $\nu_2 = N-k$

Note: k is the number of samples.

The F^* ratio (observed variance ratio) is formed by dividing the two 'mean square errors' appearing in the fourth column of the Analysis of Variance Table. In this numerical example the Analysis of Variance Table is as shown in table 8.3.

Table 8.3

Source of variation	Sum of squares	Degrees of freedom	Mean square	F^*
Between	860	$(3-1) = 2$	$\frac{860}{2} = 430$	$F^* = \frac{430}{4.37} = 98.4$
Within	118	$(30-3) = 27$	$\frac{118}{27} = 4.37$	
Total	978	$(30-1) = 29$		From F Table $F_{0.05} = 3.35$ with $\nu_1 = 2$, $\nu_2 = 27$

Note. The above discussion is a simple introduction to ANOVA. This technique has been extended to examples involving two-way classification of variables and to other more complex experimental designs. The interested reader is referred to Yamane, *Statistics*, for a detailed treatment of the ANOVA.

8.2. REGRESSION ANALYSIS AND ANALYSIS OF VARIANCE

To illustrate the similarities between regression analysis and the analysis of variance method we will work out the above example with the method of least squares regression and we will subsequently compare the results.

Let us quantify the octane-rating of the three brands of petrol, by treating their octane rating as a variable rather than as a qualitative attribute. Assume that Brand A is rated at 90 octane per gallon, Brand B at 95 octane and Brand C at 100 octane. We thus obtain a sample of 30 observations on mileage per gallon and the octane rating, which are shown in table 8.4.

Using the data of table 8.4 we obtain the regression

$$\hat{Y} = -84.5 + 1.30X_1$$

where Y = mileage per gallon
X_1 = octane rating.

To appraise these findings we need to find the correlation coefficient $R^2_{YX_1}$, and the standard errors of the parameters.

(a) From the regression results we obtain the estimate $\hat{\sigma}_u^2$:

$$\hat{\sigma}_u^2 = \frac{\Sigma e^2}{n-K} = \frac{133}{30-2} = 4.75$$

Table 8.4

N	Mileage per gallon Y_i	Octane rating X_{1i}	y_i $(Y_i - \bar{Y})$	x_{1i} $(X_{1i} - \bar{X}_1)$	$y_i x_{1i}$	y_i^2	x_{1i}^2
1	32	90	−7	−5	35	49	25
2	30	90	−9	−5	45	81	25
3	35	90	−4	−5	20	16	25
4	33	90	−6	−5	30	36	25
5	35	90	−4	−5	20	16	25
6	34	90	−5	−5	25	25	25
7	29	90	−10	−5	50	100	25
8	32	90	−7	−5	35	49	25
9	36	90	−3	−5	15	9	25
10	34	90	−5	−5	25	25	25
11	35	95	−4	0	0	16	0
12	38	95	−1	0	0	1	0
13	37	95	−2	0	0	4	0
14	40	95	1	0	0	1	0
15	41	95	2	0	0	4	0
16	35	95	−4	0	0	16	0
17	37	95	−2	0	0	4	0
18	41	95	2	0	0	4	0
19	36	95	−3	0	0	9	0
20	40	95	1	0	0	1	0
21	44	100	5	5	25	25	25
22	46	100	7	5	35	49	25
23	47	100	8	5	40	64	25
24	47	100	8	5	40	64	25
25	46	100	7	5	35	49	25
26	43	100	4	5	20	16	25
27	47	100	8	5	40	64	25
28	45	100	6	5	30	36	25
29	48	100	9	5	45	81	25
30	47	100	8	5	40	64	25
$N = 30$	$\bar{Y} = 39$	$\bar{X}_1 = 95$	$\Sigma y_i = 0$	$\Sigma x_{1i} = 0$	$\Sigma y_i x_{1i} = 650$	$\Sigma y_i^2 = 978$	$\Sigma x_{1i}^2 = 500$

(b) The variance of $\hat{b}_1$ is

$$\text{var}(\hat{b}_1) = \hat{\sigma}_u^2 \frac{1}{\Sigma x^2} = 4{\cdot}75 \frac{1}{500} = 0{\cdot}0095$$

From the variance of $\hat{b}_1$ we may compute its standard error

$$s_{(\hat{b}_1)} = \sqrt{0{\cdot}0095} \approx 0{\cdot}097$$

d the t statistic

$$t^* = \frac{\hat{b}_1}{s_{(\hat{b}_1)}} = \frac{1{\cdot}3}{0{\cdot}097} \approx 13{\cdot}3$$

(c) The correlation coefficient for the regression is

$$r^2 = 1 - \frac{\Sigma e^2}{\Sigma y^2} = 1 - \frac{133}{978} = 0{\cdot}864$$

In summary the results of the regression are

$$\hat{Y} = -84{\cdot}5 + \underset{(0{\cdot}097)}{1{\cdot}30}\ X_1$$

$$r^2 = 0{\cdot}864 \qquad \Sigma y^2 = 978 \qquad \Sigma \hat{y}^2 = 845 \qquad \Sigma e^2 = 133$$

We established in Chapter 5 that the total variation Σy^2 is split into two additive components, one component is the variation in Y explained by the regressor X_1, and the other is the unexplained variation:

$$\Sigma y^2 = \Sigma \hat{y}^2 + \Sigma e^2$$

In our example $978 = 845 + 133$

$$\begin{bmatrix}\text{Total}\\ \text{variation}\end{bmatrix} = \begin{bmatrix}\text{Explained}\\ \text{by } X_1\end{bmatrix} + \begin{bmatrix}\text{Unexplained}\\ \text{variation}\end{bmatrix}$$

This suggests that we can compile an analysis of variance table for the above regression, and use the F^* ratio to judge the overall significance of the results.

Table 8.5. Analysis of Variance Table for the Regression

Source of variation	Sum of squares	Degrees of freedom	MSE Mean Square Error	F^*
X_1	$\Sigma \hat{y}^2 = 845$	$K - 1 = 1$	$\frac{845}{1} = 845$	
Residual	$\Sigma e^2 = 133$	$N - K = 28$	$\frac{133}{28} = 4{\cdot}75$	$\frac{845}{4{\cdot}75} = 178$
Total	$\Sigma y^2 = 978$	$N - 1 = 29$		$F_{0{\cdot}05} = 4{\cdot}20$ with $\nu_1 = 1$, $\nu_2 = 28$

The observed F^* ratio is compared with the theoretical F value with $\nu_1 = K - 1 = 1$ and $\nu_2 = N - K = 28$ degrees of freedom (at the 95 per cent level of significance). From the F-Tables we find $F_{0{\cdot}05} = 4{\cdot}20$. Given that $F^* > F_{0{\cdot}05}$ we reject the null hypothesis and we accept that the regression is significant, that is X_1 is a significant explanatory factor of the variation in Y.

8.3. COMPARISON OF REGRESSION ANALYSIS AND ANALYSIS OF VARIANCE

Comparing the results of regression analysis with the results of the analysis of variance method we may draw the following conclusions.

Firstly. In both methods the total variation in Y is split into two additive components:

(a) Regression analysis

$$\sum y^2 = \sum \hat{y}^2 + \sum e^2$$

$$\text{Total} = \begin{bmatrix}\text{Explained by}\\ \text{regressor(s)}\end{bmatrix} + \begin{bmatrix}\text{Unexplained}\\ \text{(or Residual)}\end{bmatrix}$$

$$978 = 845 + 133$$

(b) Analysis of variance

$$\sum_j^k \sum_i^{n_j} (Y_{ji} - \bar{Y})^2 = \sum_j^k n_j(\bar{Y}_j - \bar{Y})^2 + \sum_j^k \sum_i^{n_j} (Y_{ji} - \bar{Y}_j)^2$$

$$\text{Total} = \text{Between} + \text{Within}$$

$$978 = 860 + 118$$

The total variation is the same in both methods. In regression analysis the data are not grouped in sub-groups or sub-samples. In the analysis of variance method the values of Y_i are grouped into sub-samples according to the values of the X_i variable.

The 'explained variation' of regression analysis corresponds to the 'between means' variation of the analysis of variance method.

The unexplained or residual variation of regression analysis corresponds to the 'within variation' of the analysis of variance approach.

Secondly. The test performed in the method of analysis of variance concerns the equality between means of sub-groups or sub-samples of an enlarged population. That is, the null hypothesis being tested is

$$H_0: \mu_1 = \mu_2 = \ldots = \mu$$

and the alternative hypothesis is

$$H_1: \mu_j \text{ not all equal}$$

The test performed in regression analysis is a test concerning the overall explanatory power of the regression as measured by R^2. The F^* ratio is a test of significance of R^2, since (as we will presently show)

$$F^* = \frac{\sum \hat{y}^2/(K-1)}{\sum e^2/(N-K)} = \frac{R^2_{YX_1}/(K-1)}{(1-R^2_{YX_1})/(N-K)}$$

If R^2 is found statistically not significant, this implies that there is no linear relationship between Y and X, that is, the true b's are zero: the null and alternative hypotheses in regression analysis are

$$H_0: b_1 = 0$$
$$H_1: b_1 \neq 0$$

Thirdly. In both methods we obtain an analysis of variance table, from which we may compute F ratios and use them for testing hypotheses related to the aim of the study.

Fourthly. It can be proved that for individual regression coefficients the t and F tests are formally equivalent, the relationship between them being

$$t^2 = F.$$

Proof. We will prove this relationship for the simple model $Y = f(X)$.

(a) Given

$$F = \frac{\sum \hat{y}^2/(K-1)}{\sum e^2/(N-K)}$$

(b) In the simple model which contains only one explanatory variable $(K-1) = 1$.

(c) We have established (in Chapter 4) that

$$\hat{y} = \hat{b}_1 x$$

Squaring through and summing over all observations we find

$$\sum \hat{y}^2 = \hat{b}_1^2 \sum x^2$$

(d) Substituting in the F ratio

$$F = \frac{\sum \hat{y}^2}{\sum e^2/(N-K)} = \frac{\hat{b}_1^2 \sum x^2}{\sum e^2/(N-K)}$$

(e) We found (in Chapter 5) that

$$t = \frac{\hat{b}_1}{s(\hat{b}_1)}$$

But

$$s(\hat{b}_1) = \sqrt{\operatorname{var}(\hat{b}_1)} = \sqrt{\sigma_u^2 \frac{1}{\sum x^2}} = \sqrt{\left[\frac{\sum e^2}{N-K}\right]\left[\frac{1}{\sum x^2}\right]}$$

Substituting in t and squaring we find

$$t^2 = \frac{\hat{b}_1^2}{[\sum e^2/(N-K)](1/\sum x^2)} = \frac{\hat{b}_1^2 \sum x^2}{\sum e^2/(N-K)} = F$$

Fifthly. Regression analysis is a more powerful method than the analysis of variance method when studying economic relationships from market data which are not experimental. Regression analysis gives all the information which we may obtain from the method of analysis of variance, but furthermore it provides numerical estimates for the influence of each explanatory variable. The analysis of variance approach shows only the addition to the explanation of total variation which one obtains by the introduction of an additional variable in the relationship. This is only part of the information provided by regression analysis as we will presently see.

It is often argued that the analysis of variance method is more appropriate for the study of the influence of qualitative factors on a certain variable.[1] This is so, the argument runs, because qualitative variables (for example profession, sex, religion) do not have numerical values, and hence their influence cannot be

[1] See K. Fox, *Intermediate Economic Statistics*, Wiley, New York, 1968, chapter 13.

assessed by regression analysis, while the analysis of variance technique does not require knowledge of the values of X's but it is based solely on the values of Y. This argument has lost a lot of its power with the expansion of the use of dummy variables in regression analysis. In most cases qualitative variables may be meaningfully approximated with dummy variables and then influence can be measured with regression analysis (see Chapter 12). The analysis of variance is most powerful (i) for the analysis of qualitative variables which cannot be meaningfully approximated by a dummy variable, and (ii) for the analysis of experimental data, where the design of the experiment permits the logical evaluation of the effects of each additional variable by determining the order in which each X is permitted to influence the value of Y. For the analysis of non-experimental data, like the data with which economists work, regression analysis is a more flexible and powerful technique (See, for example, Yamane, *Statistics*, p. 805.) However, the analysis of variance technique may be incorporated into regression analysis for carrying out tests of various hypotheses (see below).

8.4. TESTING THE OVERALL SIGNIFICANCE OF A REGRESSION

This test has been explained in the preceding section for the simple regression model including one regressor. In this section we generalise the test for models including any number of explanatory variables.

The test aims at finding out whether the explanatory variables ($X_1, X_2, \ldots, X_k$) do actually have any significant influence on the dependent variable. Formally the test of the overall significance of the regression implies testing the null hypothesis

$$H_0: b_1 = b_2 = \ldots = b_k = 0$$

against the alternative hypothesis

$$H_1: \text{not all } b_i\text{'s are zero}$$

If the null hypothesis is true, that is if all the true parameters are zero, there is no linear relationship between Y and the regressors.

The test of the overall significance may be carried out with the table of the analysis of variance. We compute the regression of Y on all the X's together and we estimate

(a) the total sum of squared deviations of the y's, Σy^2;

(b) the sum of squared deviations explained by all the regressors together, $\Sigma \hat{y}^2$;

(c) the sum of residual deviations, Σe^2.

From these terms we can evaluate the expression $\Sigma y^2 = \Sigma \hat{y}^2 + \Sigma e^2$. We next find the degrees of freedom for each of the terms of the identity. The degrees of freedom for $\Sigma \hat{y}^2$ is $K-1$, where $K(=k+1)$ is the total number of b's, including the constant intercept. The degrees of freedom for Σe^2 is $N-K$, where N is the sample size. Finally, the degrees of freedom of the total sum of

squares is $(K-1)+(N-K)=N-1$. With this information we may compute the F^* ratio as

$$F^* = \frac{\Sigma \hat{y}^2/(K-1)}{\Sigma e^2/(N-K)}$$

which is compared with the theoretical F (at the chosen level of significance) with $\nu_1 = K-1$ and $\nu_2 = N-K$ degrees of freedom. If $F^* > F$ we reject the null hypothesis, i.e. we accept that the regression is significant: not all b_i's are zero. If $F^* < F$ we accept the null hypothesis, that is we accept that the overall regression is not significant.

The above information may be summarised in a Table of Analysis of Variance (table 8.6).

Table 8.6. Analysis of Variance Table for the General Regression Model $Y = f(X_1, X_2, \ldots, X_k)$

Source of variation	*Sum of squares*	*Degrees of freedom*	*Mean Square Error*	F^*
$X_1, X_2, \ldots, X_k$	$\Sigma \hat{y}^2$	$\nu_1 = K-1$	$\frac{\Sigma \hat{y}^2}{K-1}$	$\frac{\Sigma \hat{y}^2/(K-1)}{\Sigma e^2/(N-K)} = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$
Residual	Σe^2	$\nu_2 = N-K$	$\frac{\Sigma e^2}{N-K}$	
Total	Σy^2	$N-1$		F from tables, with $\nu_1 = K-1$, $\nu_2 = N-K$ degrees of freedom

It can be shown that the F ratio for the overall significance of a regression reduces to

$$F = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

where K = number of b's (including the intercept b_0).
N = number of observations in the sample.

Proof. We have established that

$$F^* = \frac{\Sigma \hat{y}^2/(K-1)}{\Sigma e^2/(N-K)}$$

We may rewrite this expression as

$$F^* = \frac{\Sigma \hat{y}^2}{\Sigma e^2} \cdot \frac{N-K}{K-1}$$

Dividing numerator and denominator by Σy^2 we obtain

$$F^* = \frac{\Sigma \hat{y}^2/\Sigma y^2}{\Sigma e^2/\Sigma y^2} \cdot \frac{N-K}{K-1}$$

But from Chapter 7 we know that

$$\frac{\Sigma \hat{y}^2}{\Sigma y^2} = R^2_{Y.X_1\ldots X_k} \quad \text{and} \quad \frac{\Sigma e^2}{\Sigma y^2} = 1 - R^2_{Y.X_1, X_2 \ldots X_k}$$

[illegible]

[illegible] powers, and gasoline will not be consumed [illegible] with the [illegible] higher octane rating [illegible]. This is confirmed by the negative sign of b_2 [illegible] the regression [illegible] significant [illegible].

[illegible] performance [illegible] of the consumption performance of [illegible] single explanatory variable X_1 (octane [illegible]) we will gradually introduce two additional variables in the model. [illegible] weather conditions during the period of using the three [illegible] (through rain, snow, etc.) do affect the consumption of petrol irrespective of the octane rating of the petrol: in bad weather we need more petrol than in good weather, no matter whether we use high octane or low-octane petrol. We will measure 'weather' with an index of rainfall during the days of our experiment. Clearly we expect the sign of the coefficient of the weather variable b_2 to be negative. As a third explanatory variable of the petrol consumption we may consider the road and traffic conditions during our experiment. If we run the car in hilly areas with bumpy roads, we will consume, in general, more petrol than if we were driving on good roads. We will measure this factor with a dummy variable, which will assume the value of one for 'good' road conditions and the value of zero for 'bad' road conditions. We expect on a priori grounds the coefficient of this dummy variable, b_3, to appear with a positive value in the regression. The data for the regression are included in table 8.?

If we fit the simple regression

$$Y = b_0 + b_1 X_1 + u$$

we obtain the following results

$$\hat{Y} = -84.50 + 1.30\,X_1$$
$$(9.27) \quad (0.097)$$

$$R^2_{Y.X_1} = 0.864 \qquad \Sigma\hat{y}^2 = 845 \qquad \Sigma e^2 = 133$$

Table 8.? Data for the regression model

N	Mileage Y	Octane rating X_1	Weather index X_2	Road conditions X_3	y_i	x_{1i}	x_{2i}	x_{3i}	y^2
1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
10	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	35	95	101	0	−4	0	5	0·5	16
12	38	95	93	0	−1	0	−3	0·5	1
13	37	95	91	0	−2	0	−5	0·5	4
14	40	95	89	0	1	0	−7	0·5	1
15	41	95	88	0	2	0	−8	0·5	4
16	35	95	101	1	−4	0	5	0·5	16
17	37	95	97	1	−2	0	1	0·5	4
18	41	95	91	1	2	0	−5	0·5	4
19	36	95	98	1	−3	0	0	0·5	9
20	40	95	91	1	1	0	−5	0·5	1
21	44	100	91	0	5	5	−5	−0·5	25
22	46	100	93	0	7	5	−3	−0·5	49
23	47	100	96	0	8	5	0	−0·5	64
24	47	100	91	0	8	5	−5	−0·5	64
25	46	100	94	0	7	5	−2	−0·5	49
26	43	100	93	1	4	5	−3	0·5	16
27	47	100	91	1	8	5	−5	0·5	64
28	45	100	91	1	6	5	−5	0·5	36
29	48	100	89	1	9	5	−7	0·5	81
30	47	100	91	1	8	5	−5	0·5	64
$N = 30$	$\Sigma Y_i = 1{,}170$	$\Sigma X_1 = 2{,}850$	$\Sigma X_2 = 2{,}880$	$\Sigma X_3 = 15$	$\Sigma y_i = 0$	$\Sigma x_{1i} = 0$	$\Sigma x_{2i} = 0$	$\Sigma x_{3i} = 0$	$\Sigma y^2 = 978$

$\bar{Y} = 39 \quad \bar{X}_1 = 95 \quad \bar{X}_2 = 96 \quad \bar{X}_3 = 0{\cdot}5$

Table 8.7 (*cont.*)

N	x_{1i}^2	x_{2i}^2	x_{3i}^2	y_ix_{1i}	y_ix_{2i}	y_ix_{3i}	$x_{1i}x_{2i}$	$x_{1i}x_{3i}$	$x_{2i}x_{3i}$
1	25	16	0·25	35	−28	3·5	−20	2·5	−2·0
2	25	64	0·25	45	−72	4·5	−40	2·5	−4·0
3	25	36	0·25	20	−24	2·0	−30	2·5	−3·0
4	25	64	0·25	30	−48	3·0	−40	2·5	−4·0
5	25	0	0·25	20	0	2·0	0	2·5	0
6	25	0	0·25	25	0	−2·5	0	−2·5	0
7	25	196	0·25	50	−140	−5·0	−70	−2·5	7·0
8	25	81	0·25	35	−63	−3·5	−45	−2·5	4·5
9	25	49	0·25	15	−21	−1·5	−35	−2·5	3·5
10	25	36	0·25	25	−30	−2·5	−30	−2·5	3·0
11	0	25	0·25	0	−20	2·0	0	0	−2·5
12	0	9	0·25	0	3	0·5	0	0	1·5
13	0	25	0·25	0	10	1·0	0	0	2·5
14	0	49	0·25	0	−7	−0·5	0	0	3·5
15	0	64	0·25	0	−16	−1·0	0	0	4·0
16	0	25	0·25	0	−20	−2·0	0	0	2·5
17	0	1	0·25	0	−2	−1·0	0	0	0·5
18	0	25	0·25	0	−10	1·0	0	0	−2·5
19	0	0	0·25	0	0	−1·5	0	0	0
20	0	25	0·25	0	−5	0·5	0	0	−2·5
21	25	25	0·25	25	−25	−2·5	−25	−2·5	2·5
22	25	9	0·25	35	−21	−3·5	−15	−2·5	1·5
23	25	0	0·25	40	0	−4·0	0	−2·5	0
24	25	25	0·25	40	−40	−4·0	−25	−2·5	2·5
25	25	4	0·25	35	−14	−3·5	−10	−2·5	1·0
26	25	9	0·25	20	−12	2·0	−15	2·5	−1·5
27	25	25	0·25	40	−40	4·0	−25	2·5	−2·5
28	25	25	0·25	30	−30	3·0	−25	2·5	−2·5
29	25	49	0·25	45	−63	4·5	−35	2·5	−3·5
30	25	25	0·25	40	−40	4·0	−25	2·5	−2·5
$N = 30$	$\Sigma x_1^2 =$ 500	$\Sigma x_2^2 =$ 986	$\Sigma x_3^2 =$ 7·50	$\Sigma yx_1 =$ 650	$\Sigma yx_2 =$ −778	$\Sigma yx_3 =$ −1·0	$\Sigma x_1x_2 =$ −510	$\Sigma x_1x_3 =$ 0	$\Sigma x_2x_3 =$ 7·0

The simple regression of Y on X_1 explains 86 per cent of the total variation in Y, while 14 per cent remains unexplained.

If we introduce X_2 (weather) in the function, we obtain an improvement in the fit as shown by the following results.

$$\hat{Y} = -36{\cdot}88 + 1{\cdot}05\,X_1 - 0{\cdot}25\,X_2$$
$$(19{\cdot}48) \quad (0{\cdot}13) \quad (0{\cdot}09)$$

$$R^2_{Y.X_1X_2} = 0{\cdot}893 \qquad \Sigma\hat{y}^2 = 873 \qquad \Sigma e^2 = 105$$

The standard errors suggest that both variables are significant in explaining the variation in Y. Both coefficients have the correct sign, and the regression explains 89 per cent of the total variation in Y. By introducing X_2 we have managed to explain a higher proportion of the total variation in Y. We want to know whether this improvement in fit is statistically significant.

If we look at the overall significance of the two regressions we see that they both pass the F test, developed in the previous section. Thus the analysis of variance tables for these regressions are as shown in tables 8.8 and 8.9.

Table 8.8. Table of Analysis of Variance for the simple model $Y = f(X_1)$

Source	*Sum of squares*	*Degrees of freedom*	*MSE*	F^*
X_1 Residual	$\Sigma\hat{y}^2 = 845$ $\Sigma e^2 = 133$	$K - 1 = 1$ $N - K = 28$	$845/1 = 845$ $133/(N-K) = 4{\cdot}75$	$\frac{845}{4{\cdot}75} = 177{\cdot}8$
Total	$\Sigma y^2 = 978$	$N - 1 = 29$		$v_1 = 1$ $v_2 = 28$ $F_{0{\cdot}05} = 4{\cdot}20$

Table 8.9. Table of Analysis of Variance for the model $Y = f(X_1, X_2)$

Source	*Sum of squares*	*Degrees of freedom*	*MSE*	F^*
X_1, X_2 Residual	$\Sigma\hat{y}^2 = 873$ $\Sigma e^2 = 105$	$K - 1 = 2$ $N - K = 27$	$873/2 = 436$ $105/27 = 3{\cdot}9$	$\frac{436}{3{\cdot}9} = 112{\cdot}7$
Total	$\Sigma y^2 = 978$	$N - 1 = 29$		$v_1 = 2$ $v_2 = 27$ $F_{0{\cdot}05} = 3{\cdot}35$

Since in both cases $F^* > F_{0{\cdot}05}$ we conclude that both regressions are significant. However, this test of the overall significance is not very relevant for, if the regression $Y = f(X_1)$ proves to be significant, so will any relationship including X_1 and other additional variables. What we want to know is whether the new regressor X_2 has *significantly* improved the explanation in the variation of Y, in other words whether it has significantly increased the proportion of the variation explained by the first regression. For this purpose we compile another analysis of variance table as follows.

(a) From the simple regression, $Y = f(X_1)$, we obtained

$$\Sigma\hat{y}^2 = 845 \qquad \Sigma e_1^2 = 133$$

(b) From the second regression, $Y = f(X_1, X_2)$, we found

$$\Sigma\hat{y}^2 = 873 \qquad \Sigma e_2^2 = 105$$

(c) Clearly, the additional variation accounted for by the second variable X_2 is the difference

$$\Sigma\hat{y}^2 - \Sigma\hat{y}^2 = 873 - 845 = 28$$

With this information we proceed to form the analysis of variance table 8.10.

Table 8.10

Source of variation	Sum of squares	Degrees of freedom	MSE	F^*
X_1	$\Sigma\hat{y}^2 = 845$	$M-1=2-1=1$		
X_1 and X_2	$\Sigma\hat{y}^2 = 873$	$K-1=3-1=2$		
Additional variation from X_2	$\Sigma\hat{y}^2 - \Sigma\hat{y}^2 = 28$	$K-M=3-2=1$	$\frac{\Sigma\hat{y}^2 - \Sigma\hat{y}^2}{K-M} = 28$	$\frac{(\Sigma\hat{y}^2 - \Sigma\hat{y}^2)/(K-M)}{\Sigma e^2/(N-K)}$
Residual variation from $Y = f(X_1 X_2)$	$\Sigma e^2 = 105$	$N-K=30-3=27$	$\Sigma e^2/(N-K) = \frac{105}{27} = 3{\cdot}9$	$F^* = \frac{28}{3{\cdot}9} = 7{\cdot}18$
Total variation	$\Sigma y^2 = 978$			$F_{0.05} = 4{\cdot}21$, $\nu_1 = 1$, $\nu_2 = 27$

Note: M = number of all b's in the first regression (including b_0)
K = number of all b's in complete regression (including b_0).

Since F^* is greater than $F_{0.05}$ we may conclude that $b_2 \neq 0$.

It has been shown that the F test is formally equivalent to the t test which we used earlier to test the significance of b_i. On page 155 we established that $F = t^2$ for individual coefficients.

The procedure for assessing the effect of a third explanatory variable may be handled in the same way. We will present the results schematically.

The results of the regression $Y = f(X_1, X_2)$ are

$$\hat{Y} = -36{\cdot}88 + 1{\cdot}05\,X_1 - 0{\cdot}25\,X_2 \qquad R^2_{Y.X_1X_2} = 0{\cdot}893$$
$$(19{\cdot}48)\quad (0{\cdot}13)\quad (0{\cdot}09)$$

with

$$\Sigma\hat{y}^2 = 873 \qquad \Sigma e_1^2 = 105$$

We next compute the regression $Y = f(X_1, X_2, X_3)$:

$$\hat{Y} = -36{\cdot}64 + 1{\cdot}05\,X_1 - 0{\cdot}25\,X_2 + 0{\cdot}10\,X_3 \qquad R^2_{Y.X_1X_2X_3} = 0{\cdot}893$$
$$(19{\cdot}93)\quad (0{\cdot}13)\quad (0{\cdot}09)\quad (0{\cdot}74)$$

with

$$\Sigma\hat{\hat{y}}^2 = 874 \qquad \Sigma e^2 = 104$$

Hence the effect of adding X_3 is found by

$$\Sigma\hat{\hat{y}}^2 - \Sigma\hat{y}^2 = 874 - 873 = 1$$

Table 8.11 is the analysis of variance table.

Table 8.11

Source of variation	Sum of squares	Degrees of freedom	MSE	F^*
X_1, X_2	$\Sigma\hat{y}^2 = 873$	$M-1=3-1=2$		
X_1, X_2, X_3	$\Sigma\hat{\hat{y}}^2 = 874$	$K-1=4-1=3$		
Additional X_3	$\Sigma\hat{\hat{y}}^2 - \Sigma\hat{y}^2 = 1$	$(K-1)-(M-1)=3-2=1$	$1/1 = 1$	$F^* = 0{\cdot}25$
Residual from $Y = f(X_1 X_2 X_3)$	$\Sigma e^2 = 104$	$N-K=30-4=26$	$\frac{104}{26} = 4$	
Total	$\Sigma y^2 = 978$	$N-1=29$		$F_{0.05} = 4{\cdot}23$, $\nu_1 = 1$, $\nu_2 = 26$

Note: M = number of parameters in the first regression $Y = f(X_1, X_2)$
K = number of parameters in second regression $Y = f(X_1, X_2, X_3)$

The null hypothesis we are testing is $b_3 = 0$ against the alternative hypothesis: $b_3 \neq 0$.

If $F^* > F_{0.05}$ (with $\nu_1 = 1$ and $\nu_2 = 26$ degrees of freedom) we reject the null hypothesis and we accept that the third variable is an important explanatory variable. In our example $F^* < F_{0.05}$; hence we accept the null hypothesis: X_3 is not a significant variable. This is the same result as the one we reached with the standard error test.

8.5.1. GENERALISATION TO A MODEL WITH k EXPLANATORY VARIABLES

Suppose we have the model

$$Y = f(X_1, X_2 \ldots X_m, X_{m+1} \ldots X_k)$$

We first regress Y on the m variables $(X_1 \ldots X_m)$ obtaining

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \ldots + \hat{b}_m X_m$$

with $\Sigma\hat{y}^2$ and Σe_1^2 measuring the explained and unexplained parts of the total variation in y respectively.

[illegible] we introduce in the function the remaining explanatory variables $X_{m+1} \ldots X_k$ and we obtain

$$\hat{Y} = \hat{b}_0 + \ldots + \hat{b}_m X_m + \ldots + \hat{b}_k X_k$$

[illegible] $\Sigma \hat{y}^2$ and Σe^2 measuring the explained and unexplained variation. The analysis of variance table may be formed as table 8.12.

Table 8.12 Analysis of Variance Table for the General Regression model

Source of variation	Sum of squares	Degrees of freedom	MSS	
Variation in $X_1 \ldots X_m$	$\Sigma \hat{y}^{*2}$	$M-1$	$\Sigma \hat{y}^{*2}/(M-1)$	(1)
Variation in all X's $X_1 \ldots X_m \ldots X_k$	$\Sigma \hat{y}^2$	$K-1$	$\Sigma \hat{y}^2/(K-1)$	(2)
Additional variation from $X_{m+1} \ldots X_k$	$\Sigma \hat{y}^2 - \Sigma \hat{y}^{*2}$	$K-M$	$(\Sigma \hat{y}^2 - \Sigma \hat{y}^{*2})/(K-M)$	(3)
Residual variation from $Y = f(X_1 \ldots X_k)$	Σe^2	$N-K$	$\Sigma e^2/(N-K)$	(4)
Total variation in Y	Σy^2	$N-1$		

Note: $M = (m+1)$ = number of all parameters in first regression, including b_0.
$K = (k+1)$ = number of all parameters in second regression, including b_0.

With the information included in the analysis of variance table we may perform the following tests.

1. Test of the overall significance of the regression including all the k variables. The relevant F^* ratio is

$$F^* = \frac{\Sigma \hat{y}^2/(K-1)}{\Sigma e^2/(N-K)} = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

which is compared with $F_{0.05}$ with $\nu_1 = (K-1)$ and $\nu_2 = (N-K)$ degrees of freedom (for a test conducted at the 5 per cent level).

2. Test of the improvement in fit from additional regressors $X_{m+1} \ldots X_k$. The relevant F^* ratio is

$$F^* = \frac{(\Sigma \hat{y}^2 - \Sigma \hat{y}^{*2})/(K-M)}{\Sigma e^2/(N-K)}$$

which again is compared to $F_{0.05}$ with $\nu_1 = (K-M)$ and $\nu_2 = (N-K)$ degrees of freedom (for a test conducted at the 5 per cent level).

8.6. TEST OF EQUALITY BETWEEN COEFFICIENTS OBTAINED FROM DIFFERENT SAMPLES (THE CHOW TEST)

Suppose that we have two samples on the variables Y and X_1, the one containing n_1 observations and the other n_2 observations, and we use them separately for the estimation of the relationship between Y and X. We thus obtain two estimates of the same relationship for two different periods of time (or for two different cross-section samples)

$$\hat{Y}_1 = \hat{b}_0 + \hat{b}_1 X_1$$

and

$$\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

We want to test whether these two estimated relationships differ significantly, in which case we conclude that the relationship is changing from one sample to the other. For example suppose that we have the data on consumption and income for the period 1948–57, from which we estimate the consumption function

$$\hat{C}_1 = \hat{b}_0 + \hat{b}_1 Y$$

Subsequently we obtain a sample for the period 1958–67, from which we obtain the consumption function

$$\hat{C}_2 = \hat{\beta}_0 + \hat{\beta}_1 Y$$

Are the two estimated functions significantly different? Does the consumption function shift over time ($b_0 \neq \beta_0$)? Does the marginal propensity to consume (MPC) change over time ($b_1 \neq \beta_1$)? Or, is the difference insignificant, so that it may be attributed to chance, in which case we may conclude that the consumption function is stable over time?

To answer these questions we may perform the following F test suggested by Chow. (G. C. Chow, 'Tests of Equality Between Sets of Coefficients in Two Linear Regressions', *Econometrica*, vol. 28, 1960, pp. 591–605.)

Step 1. We pool together the two samples, thus forming a sample of $(n_1 + n_2)$ observations. From this we compute a 'pooled' function

$$\hat{Y}_p = \hat{a}_0 + \hat{a}_1 X$$

and we estimate the unexplained variation

$$\Sigma e_p^2 = \Sigma y_p^2 - \Sigma \hat{y}_p^2$$

with $(n_1 + n_2 - K)$ degrees of freedom. (p stands for 'pooled' and K is the total number of b's, including the intercept b_0; in our example $K = 2$.)

Step 2. We perform regression analysis on each sample separately.

From the first sample we have:

$$\hat{Y}_1 = \hat{b}_0 + \hat{b}_1 X$$
$$\Sigma e_1^2 = \Sigma y_1 - \Sigma \hat{y}_1^2$$

with $(n_1 - K)$ degrees of freedom.

From the second sample we obtain

$$\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 X$$
$$\Sigma e_2^2 = \Sigma y^2 - \Sigma \hat{y}_2^2$$

with $(n_2 - K)$ degrees of freedom.

Step 3. We add together ... in a total unexplained variation

$$(\Sigma e_1^2 + \Sigma e_2^2)$$

with $(n_1 - K) + (n_2 - K) = (n_1 + n_2 - 2K)$ degrees of freedom.

Step 4. We subtract the above sum of residual variations from the 'pooled' residual variance of Step 1, and we obtain

$$\Sigma e_p^2 - (\Sigma e_1^2 + \Sigma e_2^2)$$

with $(n_1 + n_2 - K) - (n_1 + n_2 - 2K) = K$ degrees of freedom.

Step 5. We form the ratio

$$F^* = \frac{[\Sigma e_p^2 - (\Sigma e_1^2 + \Sigma e_2^2)]/K}{(\Sigma e_1^2 + \Sigma e_2^2)/(n_1 + n_2 - 2K)}$$

The null hypothesis is $b_i = \hat{\beta}_i$, that is, there is no difference in the coefficients obtained from the two samples.

We compare the observed F^* ratio with the theoretical value of $F_{0.05}$ (or other levels of significance) with $v_1 = K$ and $v_2 = (n_1 + n_2 - 2K)$ degrees of freedom. The theoretical value of F is the value that defines the critical region of the test (at the chosen level of significance).

If $F^* > F_{0.05}$ we reject the null hypothesis, that is, we accept that the two functions differ significantly, or, the two samples give different relationships. The economic relationship being studied changes over time.

Example. Assume we have the two samples on consumption and income for the periods 1948–57 and 1958–67 which are included in table 8.13.

Table 8.13 Income and consumption data (£000 at 1958 prices)

Sample I: 1948–57			Sample II: 1958–67		
Year	Income Y_t	Consumption C_t	Year	Income Y_t	Consumption C_t
1948	17,500	12,420	1958	22,758	15,362
1949	18,253	12,690	1959	23,720	16,080
1950	18,900	13,050	1960	24,924	16,735
1951	19,126	12,863	1961	25,769	17,127
1952	19,518	12,876	1962	25,993	17,517
1953	20,413	13,450	1963	27,146	18,375
1954	21,179	13,995	1964	28,748	19,082
1955	21,911	14,559	1965	29,461	19,421
1956	22,265	14,682	1966	30,032	19,811
1957	22,706	14,985	1967	30,489	20,211

$$C = 850{\cdot}23 + 0{\cdot}63\,Y \qquad R^2_{C \cdot Y} = 0{\cdot}992$$
$$(323{\cdot}7) \quad (0{\cdot}01)$$
$$\Sigma e_p^2 = 1{,}062{,}082 = Q_1$$

with $(n_1 + n_2 - K) = 20 - 2 = 18$ degrees of freedom.

2. From the first sample the consumption function is estimated as

$$\hat{C}_1 = 3315{\cdot}27 + 0{\cdot}51\,Y \qquad R^2_{C \cdot Y} = 0{\cdot}958$$
$$(757{\cdot}7) \quad (0{\cdot}04)$$
$$\Sigma e_1^2 = 323{,}313$$

with $n_1 - K = 10 - 2 = 8$ degrees of freedom.

3. From the second sample the consumption function is found

$$\hat{C}_2 = 1545{\cdot}57 + 0{\cdot}61\,Y \qquad R^2_{C \cdot Y} = 0{\cdot}994$$
$$(421{\cdot}7) \quad (0{\cdot}01)$$
$$\Sigma e_2^2 = 128{,}552$$

with $n_2 - K = 10 - 2 = 8$ degrees of freedom.

4. The sum of the squared residuals between the two separate regressions is

$$Q_2 = \Sigma e_1^2 + \Sigma e_2^2 = 451{,}865{\cdot}4$$

with $n_1 + n_2 - 2K = 20 - 4 = 16$ degrees of freedom.

5. The difference of the above sum and the 'pooled' residuals is:

$$Q_3 = Q_1 - Q_2 = \Sigma e_p^2 - (\Sigma e_1^2 + \Sigma e_2^2) = 610{,}217$$

with $K = 2$ degrees of freedom.

6. The F^* ratio is

$$F^* = \frac{[\Sigma e_p^2 - (\Sigma e_1^2 + \Sigma e_2^2)]/K}{[\Sigma e_1^2 + \Sigma e_2^2]/(n_1 + n_2 - 2K)} = \frac{Q_3/K}{Q_2/(n_1 + n_2 - 2K)} = \frac{610{,}217/2}{451{,}865/16} = 10{\cdot}8.$$

7. The theoretical value of F at the 95 per cent level of significance with $v_1 = 2$ and $v_2 = 16$ degrees of freedom is 3·63.

Thus $F^* > F_{0 \cdot 05}$ and hence we reject the null hypothesis. The two relationships do differ significantly. That is, the consumption function changed between the two periods. Note that from the Chow test we can only infer that the function has changed. This may be due to changes in either b_0 or b_1 or both. To decide which coefficient has changed we need additional information. One way is to use dummy variables, as explained in Chapter 12. If we want to test the hypothesis that the *slope* only changes over time, we may include in the function the factor tY as an additional regressor

$$C_t = b_0 + b_1 Y_t + b_2(tY_t) + u_t$$

and test the statistical significance of $\hat{b}_2$. If $\hat{b}_2$ is found statistically significant (if $H_0: b_2 = 0$ is rejected), we may infer that the slope b changes over time, since, in this case we may write

$$C_t = \hat{b}_0 + (\hat{b}_1 + \hat{b}_2 t)Y_t$$

(See Chapter 12).

8.7. TESTING THE STABILITY OF REGRESSION COEFFICIENTS WHEN INCREASING THE SIZE OF THE SAMPLE

The aim of this test is to investigate the stability of the coefficient estimates as the sample size increases. We want to find out whether the estimates will be different in enlarged samples and whether they will remain stable over time (or in larger cross-section samples). Working with a sample a researcher may produce a regression which is too closely tailored to his sample, by experimenting with too many formulations of his model. In this case it is not certain that the estimated function will perform equally well outside the sample of data which has been used for the estimation of the coefficients. Furthermore there may have occurred events which change the structure of the relationship, for example changes in taxation laws, introduction of birth control measures, and so on. If such changes occur, the coefficients may not be stable: they may be sensitive to changes in the sample composition.

If the additional observations are more numerous than the number of parameters in the function, one may follow the procedure outlined in the previous section: that is, use the additional observations as a separate sample and apply the Chow test by computing the ratio

$$F^* = \frac{\{\Sigma e_p^2 - (\Sigma e_1^2 + \Sigma e_2^2)\}/K}{(\Sigma e_1^2 + \Sigma e_2^2)/(n_1 + n_2 - 2K)}$$

If, however, the new observations n_2 are fewer than the number of parameters in the function we may proceed as follows:

Firstly. From the augmented sample we obtain the regression

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k,$$

from which we calculate the residual sum of squares

$$\Sigma e^2 = \Sigma y^2 - \Sigma \hat{y}^2$$

with $(n_1 + n_2 - K)$ degrees of freedom.

Secondly. From the original sample of size n_1 we have

$$Y = \hat{b}_0 + \hat{b}_1 X_1 + \ldots + \hat{b}_k X_k$$

from which the unexplained sum of squares is

$$\Sigma e_1^2 = \Sigma y_1^2 - \Sigma \hat{y}_1^2$$

with $n_1 - K$ degrees of freedom.

Thirdly. Subtracting the two sums of residuals we find

$$\Sigma e^2 - \Sigma e_1^2$$

with $(n_1 + n_2 - K) - (n_1 - K) = n_2$ degrees of freedom, where n_2 are the additional observations.

Fourthly. We form the F^* ratio

$$F^* = \frac{(\Sigma e^2 - \Sigma e_1^2)/n_2}{\Sigma e_1^2/(n_1 - K)}$$

The null and alternative hypotheses are

$$H_1: b_i = \beta_i \qquad (i = 0, 1, 2, \ldots, k)$$
$$H_2: b_i \neq \beta_i$$

The F^* ratio is compared with the theoretical value of F, obtained from the F-tables with $\nu_1 = n_2$ and $\nu_2 = (n_1 - K)$ degrees of freedom.

If $F^* > F$ we reject the null hypothesis, i.e. we accept that the structural coefficients are unstable, their value changing in expanded sample periods.

Example. Suppose we have the sample of imports and income of the U.K. for the period 1950–65 as shown in table 8.14.

Table 8.14. Imports and GNP of the U.K. (in £m, at 1968 prices)

Year	*Imports (Z)*	*GNP (X)*	*Year*	*Imports (Z)*	*GNP (X)*
1950	3,748	21,777	1958	4,753	25,886
1951	4,010	22,418	1959	5,062	26,868
1952	3,711	22,308	1960	5,669	28,134
1953	4,004	23,319	1961	5,628	29,091
1954	4,151	24,180	1962	5,736	29,450
1955	4,569	24,893	1963	5,946	30,705
1956	4,582	25,310	1964	6,501	32,372
1957	4,697	25,799	1965	6,549	33,152

The import function estimated for this period (1950–65) is

$$Z = b_0 + b_1 X + u.$$

The results of the regression are

$$\hat{Z} = -2011{\cdot}85 + 0{\cdot}26\,X$$
$$(236{\cdot}71) \qquad (0{\cdot}01)$$
$$R^2 = 0{\cdot}984 \qquad \Sigma e_1^2 = 208{,}581$$

Now assume that we obtain four additional observations on imports and GNP:

	Imports	*GNP*
1966	6,705	33,764
1967	7,104	34,411
1968	7,609	35,429
1969	8,100	36,200

We want to test whether the addition of the four observations to our original sample alters significantly the coefficients of the import function.

We compute again the import equation with the enlarged sample of the twenty yearly observations

$$Z^* = -2461{\cdot}38 + 0{\cdot}28\,X$$
$$(250{\cdot}0) \qquad (0{\cdot}01)$$
$$R^2 = 0{\cdot}983 \qquad \Sigma e^2 = 573{,}069$$