Chapter 18

MSMXS2 Linear Statistical Models

(18.1) Statistical Inference

(18.1.1) Inference On The Normal Distribution

The Normal distribution is of great interest, as it can be used to model many naturally occurring events, and indeed the Central Limit Theorem means that most other distributions can be approximated by the Normal. The Normal distribution is therefore the preferred choice of distribution for modelling some population characteristic, X, say.

Suppose that $(X_1, X_2, \ldots, X_n)$ are independent and identically distributed Normal random variables, say $\mathcal{N}(\mu, \sigma^2)$. A random sample of these variables may yield data $(x_1, x_2, \ldots, x_n)$. Having observed data it is of interest as to how it can be used to predict either μ or σ or both if they are unknown.

(18.1.2) Inference When The Standard Deviation Is Known

This situation is not entirely implausible, as measuring equipment may have a known inaccuracy, giving rise to a value for the standard deviation σ. In this situation μ is estimated by the value

$$\hat{\mu} = \overline{X} = \frac{1}{n}(X_1, X_2, \ldots, X_n) \tag{1}$$

Since a sum of Normal distributions is again a Normal distribution* and so it is readily shown that $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ and so

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

This statistic may now be used for a hypothesis test, which works as follows.

- For the null hypothesis, select a suspected value of μ, say μ_0.
- For the alternative hypothesis use $\mu \neq \mu_0$.
- use the test statistic $Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ which has a standard Normal distribution when μ_0 is true.
- Reject H_0 in favour of H_1 if the value z, of Z, obtained with data $(x_1, x_2, \ldots, x_n)$ lies in the extreme of the Normal distribution, i.e. if $|z| > c$, say.

The probability

$$P = p(Z > |z| \text{ or } Z < -|z|)$$

*Although this seems 'obvious' it in fact requires rather a lot of work to show.

is called the attained significance level, or 'P value'. Typical rejection points are when P is less than 0.05, 0.01, or 0.001.

It would of course be possible to find z and then compare to the known rejection points corresponding to the required significance level and indeed this used to be standard practice. However, the invention of computers has made it possible to calculate the attained significance level, which provides information about how significant a result is.

A significance test evaluates the plausibility of a value of μ, but surely there is a whole range of values for μ that would be accepted. This is (informally) the basis of a confidence interval. A confidence interval for μ is calculated as

$$p\left(L < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < U\right) = 1 - \alpha$$

Where $1 - \alpha$ is predetermined, and L and U represent points on the standard Normal distribution that are to be determined.

It is easy to see the parallel between a confidence interval and a hypothesis test. All the values of μ that the hypothesis test accepts at the $100\alpha\%$ level lie in the $100(1-\alpha)\%$ confidence interval.

The meaning of a confidence interval is "it is $100(1-\alpha)\%$ likely that the interval will contain the true value of μ". Note that μ is fixed but unknown, while the interval is the variable quantity. Note also that in a hypothesis test, rejecting H_0 is a 'strong' result, while accepting it is weak.

(18.1.3) Inference When The Mean Is Known

Clearly, the variance will be estimated by the value

$$\hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{2}$$

However, although a sum of Normal distributions is a Normal distribution, the same does not hold for the sum of squares of a normal distribution.

Definition 3 *Suppose that $(Z_1, Z_2, \ldots, Z_n)$ are independently and identically distributed $\mathcal{N}(1,0)$. Put $Y = Z_1^2 + Z_2^2 + \cdots + Z_n^2$ then Y is said to follow a χ^2 distribution with n degrees of freedom.*

There are three main shapes to the χ^2 distribution, as are shown in Figure 1.

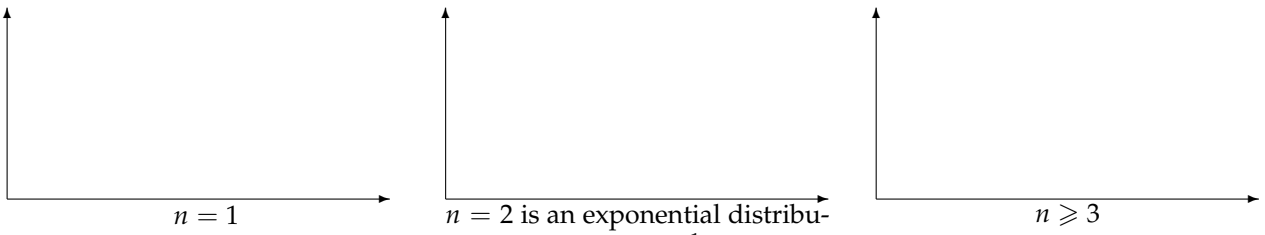

Figure 1: The three main shapes of the χ^2 distribution.

Now, $\operatorname{var} Z_i = 1$, and since $\operatorname{var} Z_i = \mathbb{E}Z_i^2 - (\mathbb{E}Z_i)^2$, it follows that $\mathbb{E}Z_i^2 = 1$. Therefore,

$$\mathbb{E}Y = \mathbb{E}\sum_{i-1}^{n} Z^2 = \sum_{i=1}^{n}\mathbb{E}Z^2 = \sum_{i=1}^{n} 1 = n$$

It can also be shown that var $Y = 2n$. Note also that

- if $Y_1 \sim \chi^2_{n_1}$ and $Y_2 \sim \chi^2_{n_2}$ are independent χ^2 distributions, then $Y_1 + Y_2 \sim \chi^2_{n_1+n_2}$.
- If $(X_1, X_2, \ldots, X_n)$ are independently and identically distributed $\mathcal{N}(\mu, \sigma^2)$ then

$$\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma}$$

are independently and identically distributed $\mathcal{N}(0,1)$ and so

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n \text{ i.e. } \sum_{i=1}^{n} (X_i - \mu)^2 \sim \sigma^2 \chi^2_n$$

When μ is known, the estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 \tag{4}$$

The sampling distribution for $\hat{\sigma}^2$ is based on the statistic

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_n \tag{5}$$

Using the χ^2 distribution is is possible to find confidence intervals and do hypothesis tests in the usual way.

The appearance of σ^2 in the statistic amy at first look a little unusual, but this is not so. The value of $\hat{\sigma}^2$ is provided by the data, making σ^2 the only variable in a similar way to the estimation of μ when σ was known.

(18.1.4) Inference When Buth Mean And Standard Deviation Are Unknown.

Perhaps the most realistic situation, it is of course the most difficult. Notice that in the estimator of σ^2 in (4) the value of μ is used. However, in this situation this must be replaced by the estimate of μ, $\overline{x}$. This produces the maximum likelihood estimator

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{x})^2$$

but this is biased. The unbiased estimator is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{x})^2 \tag{6}$$

The estimator for μ is supposed to be in standardised Normal form, but this is not possible since the variance is unknown. Notice that

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}}$$

Which is in the form $\dfrac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_n}{n}}}$ and so has a t distribution.

The estimator of σ^2 also changes, as a known value for the mean is not available. The new sampling distribution now needs to be found.

Theorem 7 *If $(Z_1, Z_2, \ldots, Z_n)$ are independently and identically distributed $\mathcal{N}(0,1)$, then* $\sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2 \sim \chi^2_{n-1}$.

This seems quite reasonable as $\sum_{i=1}^{n} (Z_i - \overline{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2$ which is the difference between a χ_n^2 and a χ_1^2 distribution. So the result is plausible.

Theorem 8 *If $(Z_1, Z_2, \ldots, Z_n)$ are independently and identically distributed $\mathcal{N}(0,1)$, then $\overline{Z}$ and $\sum_{i=1}^{n} (Z_i - \overline{Z})^2$ are independent random variables.*

Theorem 9 *If $(X_1, X_2, \ldots, X_n)$ are independently and identically distributed $\mathcal{N}(\mu, \sigma^2)$ then*

1. $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.
2. $(n-1)\frac{s^2}{\sigma^2} \sim \chi_{n-1}^2$.
3. *$\overline{X}$ and s^2 are independent random variables.*

Bearing these theorems in mind, an appropriate distribution for s^2 is now defined.

Definition 10 *If W and Z are independent random variables such that $W \sim \chi_m^2$ and $Z \sim \mathcal{N}(0,1)$ then*

$$\frac{Z}{\sqrt{\frac{W}{m}}} \sim t_m$$

is a Student's t distribution with m degrees of freedom.

Using this distribution the usual hypothesis test and confidence interval calculations can be performed using the test statistic

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \tag{11}$$

to make inference about μ.

(18.1.5) One Tailed And One Sided Tests

So far it has been assumed that in a hypothesis test the alternative hypothesis will be $\mu \neq \mu_0$. Consider the following.

Example 12 *Environmentalists[†] collect data about the levels of ozone in a forest and construct the following test.*

- H_0: $\mu = \mu_0$.
- H_1: $\mu > \mu_0$.
- *Significance level of 1%.*

The attained significance level is found to be 0.007, and on the grounds of this the environmentalists tie a small inflatable dinghy to a large supertanker.

There is a serious conceptual flaw with this test. While it is physically possible for the ozone level to be lower than anticipated, this possibility is not entertained. If it was, then the result would not have been significant. The effect of this kind of one sided test is to make results more significant than they actually are — getting the result you want.

One tailed tests, however, are not necessarily one sided, for example on the χ^2 and $\mathcal{F}$ distributions where rejecting in the lower tail would represent an exceptionally good (rather than bad) result.

[†]Cynicism suggests that environmentalists take the hypothesis testing opinion "either something is as we expect it to be, or else its worse".

(18.1.6) Comparing Two Samples

It is often of interest to compare two populations, rather than make inference about just one. In the simpler case, data values are compared and some function of each pair (difference for example) is of interest. For example, comparing measurements of the same thing taken using different equipment. The pairing of the data in this way means that the data can be treat as one sample, and a T statistic will usually be used.

The more complicated case is to determine whether two populations are 'comparable'. In hypothesis testing terms, whether or not there is a significant difference between the populations.

The $\mathcal{F}_{,}$ Distribution

The fourth major distribution used in statistical inference of this kind is the $\mathcal{F}_{,}$ distribution.

Definition 13 *If $W \sim \chi^2_m$ and $V \sim \chi^2_n$ then*

$$\frac{\frac{W}{m}}{\frac{V}{n}} \sim \mathcal{F}_{m,n}$$

which is an $\mathcal{F}_{,}$ distribution with m and n degrees of freedom.

The $\mathcal{F}_{,}$ distribution has much the same shape as the χ^2 distribution, but unlike the χ^2 which is centred around n, the $\mathcal{F}_{,}$ distribution is centred around 1.

When n is large, the $\mathcal{F}_{,}$ distribution approaches a χ^2_m. Also, $\mathcal{F}_{1,n} = t_n$.

Comparison Of Variance

When comparing means of Normally distributed samples, it has to be assumed that the variances are the same. It is therefore of interest, beforehand, to compare variances.

Suppose that

- $(X_1, X_2, \ldots, X_m)$ are independently and identically distributed $\mathcal{N}(\mu, \sigma^2)$.
- $(Y_1, Y_2, \ldots, Y_n)$ are independently and identically distributed $\mathcal{N}(\mu, \tau^2)$.

and that the Xs and the Ys are independent.

$$\text{Put} \quad S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2 \qquad \text{and} \quad S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

$$\text{Then} \quad \frac{(m-1)S_X^2}{\sigma^2} \sim \chi^2_{m-1} \qquad \text{and} \quad \frac{(n-1)S_Y^2}{\tau^2} \sim \chi^2_{n-1}$$

Hence

$$\frac{\tau^2 S_X^2}{\sigma^2 S_Y^2} \sim \mathcal{F}_{m-1,n-1}$$

This can be used to test the hypothesis $\sigma^2 = \tau^2$ against the alternative hypothesis $\sigma^2 \neq \tau^2$.

Statistical tables usually give percentage points on the $\mathcal{F}_{,}$ relating the probability to the left of some point x, so if a statistic is calculated to have value less than 1, the tables will not be of immediate use. (The $\mathcal{F}_{,}$ distribution is centred around 1). To this end, it is useful to note that

$$\mathcal{F}_{n,m} = \frac{1}{\mathcal{F}_{m,n}}$$

Comparison Of Means: Two sample t Test

It is assumed that the two Normal distributions under consideration have the same variance, hence the relevance of the previous section. Computer statistics applications can approximate the statistical test about to be described, but even so this is an unsatisfactory solution to the problem.

Suppose that

- $(X_1, X_2, \ldots, X_m)$ are independently and identically distributed $\mathcal{N}\left(\mu, \sigma^2\right)$
- $(Y_1, Y_2, \ldots, Y_n)$ are independently and identically distributed $\mathcal{N}\left(\lambda, \sigma^2\right)$

then σ^2 is estimated with the pooled sample variance,

$$\begin{aligned}\hat{\sigma}^2 = S^2 &= \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \\ &= \frac{\sum_{i=1}^{m}\left(X_i - \overline{X}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{m+n-2}\end{aligned}$$

Note that this is an unbiased estimator. S_X^2 and S_Y^2 both have a χ^2 distribution of sorts, and so it can be shown at

$$\frac{m+n-2}{\sigma^2}S^2 \sim \chi^2_{m-1} + \chi^2_{n-1} = \chi^2_{m+n-2}$$

In comparing distributions, interest lies in the 'difference' between them, and so it is found that

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu - \lambda, \frac{\sigma^2}{m} + \frac{\sigma^2}{n}\right)$$

Hence

$$\frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_n}{n}}} \quad \text{gives} \quad \frac{\frac{\overline{X}-\overline{Y}-(\mu-\lambda)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}}{\frac{1}{\sigma}\sqrt{\frac{\sum_{i=1}^{m}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{m+n-2}}} = \frac{\overline{X}-\overline{Y}-(\mu-\lambda)}{S\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t_{m+n-2}$$

Note the factor of $\dfrac{1}{\sigma}$ in the denominator. This is necessary to put the variance estimate S^2 into a form where it has a χ^2 distribution.

Using this distribution, it is possible to do the usual significance tests and confidence intervals.

This kind of two sample comparison is quite different from a paired comparison, where there is really only one sample — made by combining two other samples — under test.

It is important to perform an $\mathcal{F}$, test to ensure that the variances could be the same. Of course, this being a significance test, even if there is no significance there is no guarantee that the variances are in fact the same. Furthermore, care must be taken when working with small samples in which case the $\mathcal{F}$, test is particularly weak.

(18.2) Linear Models And Analysis Of Variance

(18.2.1) Comparison Of Means

In the previous section, two means were compared. This is now extended to k means. It is impractical to use a t test as the calculations involved would be rather excessive. Instead, two ways to estimate the variance are found, only one of which is an estimator when the means are equal. Hence an $\mathcal{F}$, test can be used.

Model 14 *With notation as described below, $y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ and are independent. This may be stated as $y_{ij} = \mu_i + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Or as $y_{ij} = \mu + \theta_i + \varepsilon_{ij}$*

The two formulations lend themselves naturally to a significance test.

With many samples being taken, the general data value will by y_{ij} which denotes the jth value from sample i. Hence the following relationships

1. $\overline{y}_i = \dfrac{1}{n_i} \sum_{j=0}^{n_i} y_{ij}$.
2. $\sum_{i=0}^{k} n_i = N$.
3. $\overline{y} = \dfrac{1}{N} \sum_{i=1}^{k} y_i = \dfrac{1}{N} \sum_{i=1}^{k} \sum_{j=0}^{n_i} y_{ij}$

Note that some texts may use a dot or bullet so show that one of the subscripts has been summed over and hence disappeared.

It is clear that a mean may be calculated for each sample, or for all of samples at once. Similarly,

- the variance of the sample means about the mean of all the samples is called the between sample variance, and is calculated by the sum

$$S_B^2 = \frac{1}{k-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\overline{y}_i - \overline{y})^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\overline{y}_i - \overline{y})^2 \tag{15}$$

- the variance for each sample can be calculated, and since the variance of a sum is the sum of variances, the within sample variance, or pooled sample variance is given by

$$S_W^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 \tag{16}$$

Now, by definition of the χ^2 variable, $\dfrac{N-k}{\sigma^2} S_W^2 \sim \chi^2_{N-k}$. Hence the between sample variance is always an estimator for σ^2, the variance of the distribution from which all the data are thought to have been drawn. It is found that this is an unbiased estimator. S_W^2 can therefore be used for inference about σ^2.

Also, when $\overline{y} = \mu$ where $\mu_i = \mu \ \forall i$ — an important null hypothesis — again by the definition of the χ^2 variable, $\dfrac{k-1}{\sigma^2} S_B^2 \sim \chi^2_{k-1}$ It is found that this is not an unbiased estimator, since

$$\mathbb{E}\, S_B^2 = \sigma^2 + \frac{\sum_{i=1}^{k} n_i (\mu_i - \tilde{\mu})^2}{k-1} \quad \text{where} \quad \tilde{\mu} = \frac{\sum_{i=1}^{k} n_i \mu_i}{N}$$

Note that μ_i is the theoretical value of $\overline{y}_i$. However, when the null hypothesis $\mu_i = \mu \ \forall i$ is true, this biasing amount will become zero, so S_B^2 is also an unbiased estimator for σ^2.

In the last few paragraphs, a hypothesis test has been repeatedly mentioned without specification. The point of this analysis is the following: It is desired to test $H_0 : \mu_i = \mu \ \forall i$ i.e. all the samples are taken from distributions which have the same mean.
From above, both S_W^2 and S_B^2 estimate σ^2 when H_0 is true. Hence the statistic $\dfrac{S_B^2}{S_W^2} \sim \mathcal{F}_{k-1,N-k}$ is of interest. Bearing in mind that an $\mathcal{F}_{,}$ has mean 1, if the value of the statistic is sufficiently greater than 1 then the null

hypothesis is rejected. If the value is less than one then S_B^2 must be exceptionally small, so the sample means are closer to μ than expected and so the null hypothesis is accepted.
This is an example of a test that is one-tailed but not one-sided.

An Analysis Of Variance

The within sample variance, S_W^2, measures the variance of the sample values around the respective group means. The between sample variance, S_B^2, measures the variance of sample means about the mean of all the data. However, since all the data is supposed to come from the same distribution, the obvious question all along has been as to the variance of all the data about the mean of all the data. This can be analysed as follows.

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i-\overline{y}+\overline{y}\right)^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{y}_i-\overline{y}\right)^2-2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)\left(\overline{y}_i-\overline{y}\right) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_i\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{y}_i-\overline{y}\right)^2 \\ &= (N-k)S_W^2+(k-1)S_B^2 \qquad (17)\end{aligned}$$

If the null hypothesis is true, then effectively there is just one large sample if size N and so

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}\right)^2 \sim \sigma^2\chi^2_{N-1}$$

This is justified (but not proved) by observing that the right hand side in (17) is a sum of two χ^2 distributions with degrees of freedom $N-k$ and $k-1$. Also, S_B^2 and S_W^2 are independent random variables.

One-Way Analysis Of Variance

Equation 17 is the basis of all analysis of variance. What the equation says is that the sum of squares can be split between variability of the group means about the actual mean, and some residual error in variability of data about their respective group means.

Implicit in this is the hypothesis that the group means are equal, and indeed if the sum of squares partitioned into S_B^2 is significantly large ($\mathcal{F}$ test) then this hypothesis should be rejected. The virtue of expressing the residual error about respective group means is that even if the null hypothesis is false S_W^2 is still a valid estimator of σ^2, the variance of all the data about is mean.

In practice it is possible that there may be many clear ways to group the data. For example, the height of children may be grouped by age, sex, weight, etc. Once the grouping has been decided upon S_B^2 and S_W^2 can be calculated, and this is a one-way analysis of variance. This is usually set out in a table, as shown theoretically in Table 1

The term "mean square" refers to the sum of squares divided by its degrees of freedom. This produces the number S_B^2 or S_W^2, the ratio of which produces the required $\mathcal{F}$ statistic.

Source	D.F.	Sum Of Squares	Mean Square	$\mathcal{F}$ statistic	'P' value
Between samples	$k-1$	$\sum_{i=1}^{k} n_i \left(\overline{y}_i - \overline{y}\right)^2$	S_B^2	$S_B^2 \div S_W^2$	$1 - p(\mathcal{F} < f)$
Within samples	$N-k$	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{y}_{ij} - \overline{y}_i\right)^2$	S_W^2		
Total	$N-1$	$\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\overline{y}_{ij} - \overline{y}\right)^2$			

Table 1: One-way analysis of variance

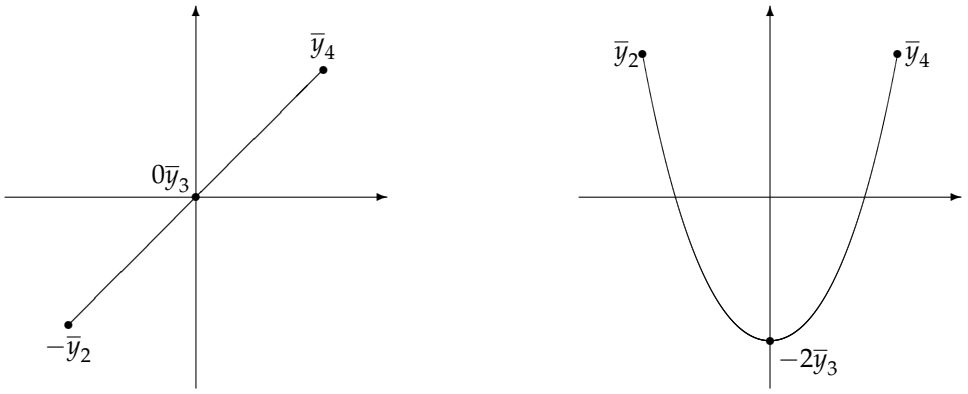

Figure 2: Interpretation of contrasts.

Orthogonal Contrasts

Having performed a one-way analysis of variance the between group sum of squares can be further partitioned to find what proportion of the sum of squares arises from which group. It could be the case that only a few of the groups are of interest rather than all of them. Any such situation can be modeled by taking a linear combination of the groups. The coefficients of such a combination form a row vector, which is called a contrast if $\sum_{i=1}^{k} l_i = 0$. Contrasts $\mathbf{l}$ and $\mathbf{l}'$ are orthogonal if $\mathbf{l} \cdot \mathbf{l}' = 0$ and are of interest as they can be used to partition the between group sum of squares.

The elements of a contrast may be chosen at will—provided of course that they add up to zero. However, what contrasts to use should be decided before receiving the data to make sure that they have a realistically meaningful interpretation. Contrasts can be used to identify relationships between the group and its mean — say the children's height increases with their weight. Provided that the values of the group are equally spaced, the contrast of the form $\begin{pmatrix} -1 & 0 & 1 \end{pmatrix}$ corresponds a linear relationship since two points supplied by the two groups can be used to determine the line, and the weights in the contrast correspond to where the line ought to be. The contrast can be of any length as long as it contains a -1 and a 1 and the other entries are zero. This is illustrated in Figure 2. Similarly, three points are needed to determine a parabola, and so a second contrast corresponds to a quadratic relationship between treatment and group mean.

The between group sum of squares can be 'resolved' in the 'direction' of the contrasts. The relationship between the elements of the contrast (the last is a linear combination of the preceding ones) effectively remove a dimension from the vector space — corresponding to a degree of freedom.

It is assumed that each of the k groups has the same number of data values in it, n. Where the contrasts are

$l_1, l_1, \ldots, l_{k-1}$ with elements denoted by l_{i_j}, the formula is

$$S_B^2 = n\frac{\left(\sum_{i=1}^k l_{1_i}\overline{y}_i\right)^2}{\sum_{i=1}^k \left(l_{1_i}\right)^2} + n\frac{\left(\sum_{i=1}^k l_{2_i}\overline{y}_i\right)^2}{\sum_{i=1}^k \left(l_{2_i}\right)^2} + \cdots + n\frac{\left(\sum_{i=1}^k l_{(k-1)_i}\overline{y}_i\right)^2}{\sum_{i=1}^k \left(l_{(l-1)_i}\right)^2}$$

When the null hypothesis—that the group means are equal—is true, each component is independent and has distribution $\sigma^2\chi_1^2$. $\mathcal{F}$ tests can be performed by finding the ratio of contrast partition mean square over error mean square.

Two Way Analysis Of Variance

It was mentioned above that there may be more than one possible grouping of the data. Each datum can then be though of as a combination of two distinct effects, and an error, as set out in Table 2.

	B_1	B_2	$\ldots$	B_b
A_1	y_{11}	y_{12}	$\ldots$	y_{1b}
A_2	y_{21}	y_{22}	$\ldots$	y_{2b}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
A_a	y_{a1}	y_{a2}	$\ldots$	y_{ab}

Table 2: Two effects contribute to the value of a datum.

Model 18 *The data are modeled by the formula*

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where α and β represent the A and B effects, and ε is the error.

To ensure the number of parameters being estimated does not get too out of hand, the constraint

$$\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0$$

is imposed, giving least squares estimates

$$\hat{\mu} = \overline{y} \qquad \hat{\alpha}_i = \overline{y}_i - \overline{y} \qquad \hat{\beta}_j = \overline{y}_j - \overline{y}$$

A very important assumption is that the two effects under consideration are independent so that the effect of A does not change with the level of B under consideration. If this was not the case then the sum of squares attributable to A would be different for each B. This assumption means that if a one way analysis has already been performed for A, then the sum of squares for it does not change. What the two way analysis does, therefore, is take some of the error and attribute it to another effect.

A one-way analysis may be readily extended to a two way one since the between groups sum of squares already identified remains the same—this is a direct result of the additivity assumption. It can be shown that

$$\sum_{i=1}^a \sum_{j=1}^b \left(\varepsilon_{ij}\right)^2 = b\sum_{i=1}^a \left(\overline{y}_i - \overline{y}\right)^2 + a\sum_{j=1}^b \left(\overline{y}_j - \overline{y}\right)^2 + \sum_{i=1}^a \sum_{j=1}^b \left(y_{ij} - \overline{y}_i - \overline{y}_j + \overline{y}\right)^2$$

If in each group there are r data, then this formula becomes

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(\varepsilon_{ijk}\right)^2 = br\sum_{i=1}^{a}(\overline{y}_i-\overline{y})^2 + ar\sum_{j=1}^{b}\left(\overline{y}_j-\overline{y}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk}-\overline{y}_i-\overline{y}_j+\overline{y}\right)^2$$

It is clear to see that this is in the form of sums of squares for the A effect, the B effect, and the residual error. A two-way analysis of variance is summarised in Table 3

Source	D.F.	Sum Of Squares	Mean Square	$\mathcal{F}$ Statistic
Effect A	$a-1$	$\sum_{i=1}^{a} b\left(\overline{y}_i-\overline{y}\right)^2$	S_A^2	$\frac{S_A^2}{S_E^2}$
Effect B	$b-1$	$\sum_{j=1}^{b} a\left(\overline{y}_j-\overline{y}\right)^2$	S_B^2	$\frac{S_B^2}{S_E^2}$
Error	$(a-1)(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\left(y_{ij}-\overline{y}_i-\overline{y}_j+\overline{y}\right)^2$	S_E^2	
Total	$ab-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\left(y_{ij}-\overline{y}\right)^2$		

Table 3: Table for two-way analysis of variance

(18.2.2) Straight Line Regression

Model 19 *For paired data (x_i, y_i) for $1 \leqslant i \leqslant n$ there exists a relationship between the paired values modeled by*

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables each with expected value 0 and variance σ^2.

It is usually assumed that the ε_is are Normally distributed so that $y_i \,|\, x_i \sim \mathcal{N}\left(\alpha+\beta x_i, \sigma^2\right)$. An estimate for α and β, $\hat{\alpha}$ and $\hat{\beta}$, may be found by a maximum likelihood estimate as follows.

Estimating α And β.

The quantity $\hat{\varepsilon} = y_i - \hat{\alpha} - \hat{\beta}x_i$ is called the ith residual, and $\hat{\alpha}$ and $\hat{\beta}$ must be found so that the deviation of the data from the line described by the model is minimal. The residual sum of squares given by $\sum_{i=1}^{n}\varepsilon_i^2$ which needs to be minimised.

$$RSS = \sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

$$\frac{\partial RSS}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i) = -2(n\overline{y}-n\alpha-n\beta\overline{x}) \tag{20}$$

$$\frac{\partial RSS}{\partial \beta} = -2\sum_{i=1}^{n}x_i(y_i-\alpha-\beta x_i) \tag{21}$$

From (20) it is evident that $\frac{\partial RSS}{\partial \alpha} = 0$ by putting $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$. Substituting this into (21) gives

$$\begin{aligned}
\frac{\partial RSS}{\partial \beta} &= -2\sum_{i=1}^{n} x_i \left(y_i - \left(\overline{y} - \hat{\beta}\overline{x}\right) - \hat{\beta}x_i\right) \\
0 &= -2\sum_{i=1}^{n} x_i y_i + 2\overline{y}\sum_{i=1}^{n} x_i - 2\hat{\beta}\overline{x}\sum_{i=1}^{n} x_i + 2\hat{\beta}\sum_{i=1}^{n} x_i^2 \\
&= -\sum_{i=1}^{n} x_i y_i + n\overline{x}\overline{y} + \hat{\beta}\left(-n\overline{x}^2 + \sum_{i=1}^{n} x_i^2\right) \\
\hat{\beta} &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}
\end{aligned}$$

At this point it is convenient to introduce a notation convention.

- Sums of squares may be expressed in the form S_{xy} where,
  - $S_{xx} = \sum x_i^2$.
  - $S_{xy} = \sum x_i y_i$.
  - $S_{yy} = \sum y_i^2$.
- Corrected sums of squares may be expressed in the form C_{xy} where,
  - $C_{xx} = \sum (x_i - \overline{x})^2$.
  - $C_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$.
  - $C_{yy} = \sum (y_i - \overline{y})^2$.

The estimate for β can be represented in the alternative forms

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{C_{xy}}{C_{xx}} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Note that the last expression is allowed since

$$\sum_{i=1}^{n} (-x_i\overline{y} + \overline{x}\overline{y}) = -n\overline{x}\overline{y} + n\overline{x}\overline{y} = 0$$

Sampling Distributions For The Estimators

Having produced estimators, it is of interest as to how they behave—their sampling distributions.

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\
&= \sum_{i=1}^{n} a_i y_i \quad \text{where } a_i = \frac{x_i-\overline{x}}{C_{xx}} \\
\mathbb{E}\,\hat{\beta} &= \sum_{i=1}^{n} a_i\,\mathbb{E}\,y_i \\
&= \sum_{i=1}^{n} a_i(\alpha+\beta x_i) \\
&= \frac{1}{C_{xx}}\left(\alpha\sum_{i=1}^{n}(x_i-\overline{x}) + \beta\sum_{i=1}^{n}x_i(x_i-\overline{x})\right) \\
&= \frac{1}{C_{xx}}(0+\beta C_{xx}) \\
&= \beta
\end{aligned}$$

So $\hat{\beta}$ is an unbiased estimator. Similarly

$$\begin{aligned}
\operatorname{var}\hat{\beta} &= \sum_{i=1}^{n} a_i^2 \operatorname{var} y_i \quad a_i = \frac{x_i-\overline{x}}{C_{xx}} \\
&= \sigma^2 \sum_{i=1}^{n} a_i^2
\end{aligned}$$

The variance of y_i is σ^2 which comes from the ε_is since $y_i = \alpha + \beta x_i + \varepsilon_i$

$$\begin{aligned}
&= \frac{\sigma^2\sum_{i=1}^{n}(x_i-\overline{x})^2}{C_{xx}^2} \\
&= \frac{\sigma^2}{C_{xx}}
\end{aligned}$$

So the mean and variance of the estimator $\hat{\beta}$ have been found. Note that this is the variance of the estimator, not an estimate for variance: It is the mean square error (MSE) as discussed in Chapter **??**.

From the model, $y_i \,|\, x_i \sim \mathcal{N}\left(\alpha+\beta x_i, \sigma^2\right)$ and since

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

it can be seen that $\hat{\beta}$ is a sum of independent Normal distributions and so itself has the Normal distribution $\hat{\beta} \sim \mathcal{N}\left(\beta, \dfrac{\sigma^2}{C_{xx}}\right)$. A similar process can now be done for $\hat{\alpha}$, but note that

$$y_i = \alpha + \beta x_i \quad \Rightarrow \quad \overline{y} = \alpha + \beta\overline{x}$$

Now, from the maximal likelihood estimation process above,

$$\begin{aligned}\hat{\alpha} &= \overline{y} - \hat{\beta}\overline{x} \\ \mathbb{E}\,\hat{\alpha} &= \overline{y} - \mathbb{E}\,\left(\hat{\beta}\overline{x}\right) \\ &= \overline{y} - \overline{x}\,\mathbb{E}\,\hat{\beta} \\ &= \overline{y} - \beta\overline{x} \\ &= \alpha\end{aligned}$$

Hence $\hat{\alpha}$ is also an unbiased estimator. For the variance,

$$\begin{aligned}\operatorname{var}\hat{\alpha} &= \operatorname{var}\left(\overline{y} - \hat{\beta}\overline{x}\right) \\ &= \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + \operatorname{var}\left(\hat{\beta}\overline{x}\right) \\ &= \frac{1}{n^2}(n\sigma^2) + \overline{x}^2\frac{\sigma^2\overline{x}^2}{C_{xx}} \\ &= \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{C_{xx}}\right)\end{aligned}$$

It is observed that $\hat{\alpha}$ is a linear combination of Normal distributions, and so $\hat{\alpha} \sim \mathcal{N}\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{C_{xx}}\right)\right)$.

The Normal distribution of $\hat{\alpha}$ and $\hat{\beta}$ is actually conditional on the ε_is being Normally distributed, which in this model they are. Furthermore $\hat{\beta}$ and $\overline{y}$ are independent random variables.

Estimating The Variance

Estimating the variance is really a question of estimating the ε_is, since they *are* the variance. Having used the data—the x_is and the y_is—to generate estimates for α and β it is now possible to work out what the model thinks the y_is should be. It is usual to think of the x_is as being fixed and the y_is depending on them. Hence

$$\hat{y}_i \stackrel{\text{def}}{=} \hat{\alpha} + \hat{\beta}x_i$$

Clearly there will be a discrepancy between what the y_is are, and what the model says they are supposed to be. This amount is called the residual,

$$\hat{\varepsilon}_i \stackrel{\text{def}}{=} y_i - \hat{\alpha} - \hat{\beta}x_i = y_i - \hat{y}_i$$

The variance is then estimated by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\hat{\varepsilon}^2}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The term in the denominator is plausible since two parameters are being estimated. If the ε_is are Normally distributed then $\frac{n-2}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-2}$.

Since the variance estimate has a χ^2 distribution and the estimators $\hat{\alpha}$ and $\hat{\beta}$ are Normally distributed, a t test can be constructed to test whether particular values of α and β are consistent with the data. Recall that

a T statistic is of the form $\dfrac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_n}{n}}}$. These will be calculated in due course, but first of all note

$$\begin{aligned} C_{xx} + n\overline{x}^2 &= \sum_{i=1}^{n} (x_i - \overline{x})^2 + n\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \overline{x}^2 + n\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 \end{aligned} \tag{22}$$

Using equation (22) to simplify the expression for the statistic for α,

$$\frac{\frac{\hat{\beta}-\beta}{\sqrt{\frac{\sigma^2}{C_{xx}}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \sim t_{n-2} \qquad \frac{\frac{\hat{\alpha}-\alpha}{\sigma\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{C_{xx}}}}}{\sqrt{\frac{\hat{s}^2}{\sigma^2}}} \sim t_{n-2}$$

$$\frac{\left(\hat{\beta}-\beta\right)\sqrt{C_{xx}}}{\hat{\sigma}} \sim t_{n-2} \qquad \frac{(\hat{\alpha}-\alpha)}{\hat{\sigma}}\sqrt{\frac{nC_{xx}}{\sum_{i=1}^{n} x_i^2}} \sim t_{n-2}$$

Hence the usual t tests can be performed to verify suspected values of α and β. In particular it is possible to test for zero slope, although equivalently an $\mathcal{F}$ test could be used on the ratio of mean squares as found by the analysis of variance.

Analysis Of Variance

As well as significance tests, the next part of the repertoire of statistical techniques is an analysis of variance. In this case the variance can be attributed to variability accounted for by the model—the regression sum of squares—and the residual sum of squares. Now,

$$\begin{aligned} \sum_{i=1}^{n} (y_i - \overline{y})^2 &= \sum_{i=1}^{n} \left(\hat{y}_i + \hat{\varepsilon}_i - \hat{\alpha} - \hat{\beta}\overline{x}\right)^2 \\ &= \sum_{i=1}^{n} \left(\hat{\alpha} + \hat{\beta}x_i + \hat{\varepsilon}_i - \hat{\alpha} - \hat{\beta}\overline{x}\right)^2 \\ &= \sum_{i=1}^{n} \left(\hat{\beta}x_i + \hat{\varepsilon}_i - \hat{\beta}\overline{x}\right)^2 \\ &= \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} \hat{\varepsilon}^2 + 2\hat{\beta}\sum_{i=1}^{n} (x_i - \overline{x})\hat{\varepsilon}_i \\ &= \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} \hat{\varepsilon}^2 \end{aligned}$$

which is the required result. The total sum of squares has been written as a 'between groups' and 'within groups' sum of squares.

(18.2.3) The General Linear Model

First of all note that the General Linear Model should not be confused with the Generalised Linear Model, which is something rather different. Recall that in the analysis of variance it was possible to attribute error in the data to one or two parameters. The General Linear Model provides a way to consider many parameters, as well as fitting a straight line or indeed a hyperplane. The General Linear Model is therefore of great

interest.

Model 23 (The General Linear Model) *Let the observed data form an $n \times 1$ column vector $\mathbf{y}$, let β be a $p \times 1$ column vector of parameters ($p < n$), and let ε be an $n \times 1$ column vector or errors. The data may then me modeled by the equation*

$$\mathbf{y} = X\beta + \varepsilon$$

where X is an $n \times p$ matrix called the design matrix.

The design matrix may consist of numbers, for example in the case of the k sample problem, as well as x_is when fitting a straight line or hyperplane. It is assumed that X is of full rank (namely p) so that it has linearly independent columns. If this is not so then it is often possible to reparameterise the model and so solve this problem. Since $X^T X$ is a $p \times p$ matrix, X being of full rank means that $X^T X$ is non-singular and so its inverse exists. As usual, each element of ε has $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$.

Estimating The Unknown Parameters

The first job with any particular application of the General Linear Model is to estimate the unknown parameters. Since $\varepsilon = \mathbf{y} - X\beta$, the residual sum of squares is given by

$$R(\hat{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2(\hat{\beta}) = (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) \tag{24}$$

The estimate of the parameters, $\hat{\beta}$, is now chosen to minimise this error.

Theorem 25 *Suppose that β_0 satisfies the equation*

$$X^T X\beta_0 = X^T\mathbf{y}$$

then $R(\hat{\beta}) \geqslant R(\beta_0)$ for all $\hat{\beta}$.

Proof. From equation (24),

$$\begin{aligned} R(\hat{\beta}) - R(\beta_0) &= (\mathbf{y} - X\hat{\beta})^T(\mathbf{y} - X\hat{\beta}) - (\mathbf{y} - X\beta_0)^T(\mathbf{y} - X\beta_0) \\ &= \left(\mathbf{y}\mathbf{y}^T - \hat{\beta}^T X^T\mathbf{y} - \mathbf{y}^T X\hat{\beta} + \hat{\beta}^T X^T X\hat{\beta}\right) - \left(\mathbf{y}\mathbf{y}^T - \beta_0^T X^T\mathbf{y} - \mathbf{y}^T X\beta_0 + \beta_0^T X^T X\beta_0\right) \end{aligned}$$

But $\beta^T X^T\mathbf{y} = \left(\mathbf{y}^T X\beta\right)^T = \mathbf{y}^T X\beta$ since it is a scalar. Hence

$$\begin{aligned} &= \left(\mathbf{y}\mathbf{y}^T - 2\hat{\beta}^T X^T\mathbf{y} + \hat{\beta}^T X^T X\hat{\beta}\right) - \left(\mathbf{y}\mathbf{y}^T - 2\beta_0^T X^T\mathbf{y} + \beta_0^T X^T X\beta_0\right) \\ &= \left(\mathbf{y}\mathbf{y}^T - 2\hat{\beta}^T X^T X\beta_0 + \hat{\beta}^T X^T X\hat{\beta}\right) - \left(\mathbf{y}\mathbf{y}^T - 2\beta_0^T X^T X\beta_0 + \beta_0^T X^T X\beta_0\right) \quad \text{by hypothesis} \\ &= \hat{\beta}^T X^T X\hat{\beta} - 2\hat{\beta}^T X^T X\beta_0 + \beta_0^T X^T X\beta_0 \end{aligned}$$

But $\hat{\beta}^T X^T X \beta_0$ is a scalar, and so is equal to its transpose. Hence

$$\begin{aligned}
&= \hat{\beta}^T X^T X \hat{\beta} - \hat{\beta}^T X^T X \beta_0 - \beta_0^T X^T X \hat{\beta} + \beta_0^T X^T X \beta_0 \\
&= (\hat{\beta}^T - \beta_0^T) X^T X (\hat{\beta} - \beta_0) \\
&= \left(X\left(\hat{\beta} - \beta_0\right)\right)^T \left(X\left(\hat{\beta} - \beta_0\right)\right) \\
&\geqslant 0 \\
&= 0 \Leftrightarrow \hat{\beta} = \beta_0
\end{aligned}$$

Hence the lemma has been shown. □

What this means is that the least squares estimate for β satisfies $X^T X \hat{\beta} = X^T \mathbf{y}$ and since $X^T X$ is non-singular this can be solved to give $\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$. This estimator is readily seen as being unbiased since

$$\mathbb{E}\,\hat{\beta} = \mathbb{E}\left((X^T X)^{-1} X^T \mathbf{y}\right) = (X^T X)^{-1} X^T \,\mathbb{E}\,\mathbf{y} = (X^T X)^{-1} X^T X \beta = \beta$$

Furthermore the variance of the estimator can be found in the following way.

$$\begin{aligned}
\operatorname{var} \hat{\theta}_i &= \operatorname{var}\left(\left(X^T X\right)^{-1} X^T\right) y_i \\
&= \operatorname{var} \sum_{j=1}^{n} \left(\left(X^T X\right)^{-1} X^T\right)_{ij} y_i \\
&= \sum_{j=1}^{n} \left(\left(X^T X\right)^{-1} X^T\right)_{ij}^2 \operatorname{var} y_i
\end{aligned}$$

Now, $\sum_j (A)_{ij}^2 = \sum_j (A)_{ij}(A)_{ij} = (AA^T)_{ii}$. Hence

$$\begin{aligned}
&= \left(\left(X^T X\right)^{-1} X^T \left(\left(X^T X\right)^{-1} X^T\right)^T\right)_{ii} \operatorname{var} y_i \\
&= \left(\left(X^T X\right)^{-1}\right)_{ii} \sigma^2 \qquad (26)
\end{aligned}$$

This result can be used to perform t test to assess the ability of individual parameters to explain variability in the data. In a similar way to fitting a straight line, the fitted values and residual errors are defined as follows.

Definition 27 *Where $H = X(X^T X)^{-1} X^T$, which is called the hat matrix,*

1. $\hat{\mathbf{y}} \stackrel{\text{def}}{=} H\mathbf{y}$.
2. $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = (I - H)\mathbf{y}$.

Theorem 28 *The hat matrix H has the following properties.*

1. $H^T = H = H^2$ *so H is symmetric and idempotent.*
2. $(I - H)^T = (I - H) = (I - H)^2$.
3. $HX = X$.
4. $\operatorname{tr} H = p$.

Proof. Taking each part in turn,

1. From the definition of the hat matrix,

$$H^T = \left(X(X^TX)^{-1}X^T\right)^T = X^{T^T}(X^TX)^{-1^T}X^T = H$$

and also

$$H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = H$$

2. Follows from 1.

3. Clearly

$$HX = X(X^TX)^{-1}X^TX = X$$

4. Since $\text{tr}\,(AB) = \text{tr}\,(BA)$,

$$\text{tr}\left(X(X^TX)^{-1}X^T\right) = \text{tr}\left((X^TX)^{-1}X^TX\right) = \text{tr}\, I = p$$ □

Estimating σ^2

The obvious way to partition the total variability experienced is into that which can be accounted for by the model, and that which cannot. Where

$$\hat{\mathbf{y}} = H\mathbf{y} \qquad \text{and} \qquad \hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$$

observe that

$$\sum_{i=1}^n \hat{y}_i^2 = \hat{\mathbf{y}}^T\hat{\mathbf{y}} = \mathbf{y}^TH^TH\mathbf{y} = \mathbf{y}^TH\mathbf{y}$$

$$\text{and} \quad \sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \mathbf{y}^T(I-H)^T(I-H)\mathbf{y} = \mathbf{y}^T(I-H)\mathbf{y}$$

$$\text{and so} \quad \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 = \mathbf{y}^TH\mathbf{y} + \mathbf{y}^T(I-H)\mathbf{y} = \sum_{i=1}^n y_i^2 = \mathbf{y}^T\mathbf{y}$$

Hence the important result $\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2$. This calculation is the basis of the analysis of variance which follows in due course, however, it does not have the usual interpretation of 'variabilities'—$\sum_{i=1}^n y_i^2$ is simply the sum of squares of the 'dependent' data. Compare the form to the equation $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$.

In the usual way, the variance estimate is formed from the sum of squares of the data values about their expected value, $\hat{\mathbf{y}}$. The estimator is therefore

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-p} \tag{29}$$

This can be shown to be unbiased. First of all note the result

$$\begin{aligned}
\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} &= \mathbf{y}^T(I-H)^T(I-H)\mathbf{y} \\
&= (X\grave{} + \grave{\varepsilon})^T(I-H)(X\grave{} + \grave{\varepsilon}) \\
&= \grave{\varepsilon}^T(I-H)\grave{\varepsilon} + (X\grave{})^T(I-H)(X\grave{}) \\
&= \grave{\varepsilon}^T(I-H)\grave{\varepsilon} - \grave{}^TX^TX\grave{} - \grave{}^TX^THX\grave{} \\
&= \grave{\varepsilon}^T(I-H)\grave{\varepsilon}
\end{aligned}$$

Now, since $\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}$ is a scalar (or 1×1 matrix), it is equal to its trace. Hence

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}\right) &= \mathbb{E}\ \mathrm{tr}\left(\left(\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}\right)\right)\\ &= \mathbb{E}\left(\mathrm{tr}\left(\boldsymbol{\varepsilon}^T(I-H)\boldsymbol{\varepsilon}\right)\right)\\ &= \mathbb{E}\left(\mathrm{tr}\left((I-H)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)\right)\\ &= \mathrm{tr}\left((I-H)\,\mathbb{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right)\right)\end{aligned}$$

But $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ is an $n \times n$ matrix, with ij entry $\varepsilon_i\varepsilon_j$. The expected value of a matrix is the matrix of expected values, and by definition $\mathbb{E}\,\varepsilon_i = 0$ and since the ε_is are independent, clearly the off diagonal entries in the matrix of expected values will all be 0. However, $\mathbb{E}\,\varepsilon_i^2 = \mathrm{var}\,\varepsilon_i - (\mathbb{E}\,\varepsilon_i)^2 = \sigma^2$. Continuing,

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}\right) &= \mathrm{tr}\left((I-H)\sigma^2 I\right)\\ &= \sigma^2\left(\mathrm{tr}\,I - \mathrm{tr}\,H\right)\\ &= \sigma^2(n-p)\end{aligned}$$

Hence it is clear that (29) is an unbiased estimator of σ^2.

Analysis Of Variance

After fitting a model, the elements of the vector $\hat{\boldsymbol{\varepsilon}}$ represent the remaining discrepancy between the model and the actual data—a residual error. Now, unlike the k sample problem the data here is supposed to vary. The total variability of the data is therefore the data itself. The equation $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ means that the data has some in-built variability, as well as the usual sampling error. Since a constant is also part of the model, it is expected that $y_i = 0\ \forall i$. The partition of the total sum of squares into the model sum of squares and residual error is therefore

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \quad \text{or equivalently} \quad \mathbf{y}^T\mathbf{y} = \hat{\mathbf{y}}^T\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}$$

Now,

$$\begin{aligned}\hat{\mathbf{y}}^T\hat{\mathbf{y}} &= (H\mathbf{y})^T(H\mathbf{y})\\ &= \left(\boldsymbol{\beta}^T X^T + \boldsymbol{\varepsilon}^T\right) H\left(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right)\\ &= (X\boldsymbol{\beta})^T X\boldsymbol{\beta} + \boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon} + 2\boldsymbol{\varepsilon}^T X\boldsymbol{\beta}\end{aligned}$$

Inkeeping with the view that the data should all be zero, the null hypothesis $\boldsymbol{\beta} = \mathbf{0}$ is under test, and if true this gives $\hat{\mathbf{y}}^T\hat{\mathbf{y}} = \boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}$ which is a scalar. Now, each y_i has a Normal distribution, and is a linear function of p parameters; it follows therefore that each parameter is Normally distributed, and that the sum of squares $\hat{\mathbf{y}}^T\hat{\mathbf{y}} = \boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}$ is therefore a sum of p squared Normal distributions. Hence $\hat{\mathbf{y}}^T\hat{\mathbf{y}} \sim \sigma^2\chi_p^2$ is the null distribution. This partition of the sum of squares is called the Model sum of squares, S_M^2. Observe that

$$\begin{aligned}\mathbb{E}\,S_M^2 &= (X\boldsymbol{\beta})^T(X\boldsymbol{\beta}) + \mathbb{E}\left(\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}\right) + 2\,\mathbb{E}\left(\boldsymbol{\varepsilon}^T X\boldsymbol{\beta}\right)\\ &= (X\boldsymbol{\beta})^T(X\boldsymbol{\beta}) + \mathbb{E}\left(\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}\right) + 2\,\mathbb{E}\left(\boldsymbol{\varepsilon}^T(\mathbf{y}-\boldsymbol{\varepsilon})\right)\\ &= (X\boldsymbol{\beta})^T(X\boldsymbol{\beta}) + \mathbb{E}\left(\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}\right) + 2\,\mathbb{E}\left(\boldsymbol{\varepsilon}^T\mathbf{y}\right) - 2\,\mathbb{E}\left(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right)\\ &= (X\boldsymbol{\beta})^T(X\boldsymbol{\beta}) + \mathbb{E}\left(\boldsymbol{\varepsilon}^T H\boldsymbol{\varepsilon}\right) \quad \text{by independence, using } \mathbb{E}\,\boldsymbol{\varepsilon} = \mathbf{0}\\ &= \left(\mathbb{E}\,\mathbf{y}\right)^T\left(\mathbb{E}\,\mathbf{y}\right) + p\sigma^2 \quad \text{since } \mathrm{tr}\,H = p.\end{aligned}$$

Source	D.F.	S.S.	Mean Square	Expected Mean Square
Fitting the model	p	$\hat{\mathbf{y}}^T\hat{\mathbf{y}}$	$\frac{\hat{\mathbf{y}}^T\hat{\mathbf{y}}}{p}$	$\frac{(X\grave{})^T(X\grave{})}{p}+\sigma^2$
Residual	$n-p$	$\hat{\boldsymbol{\prime\prime}}^T\hat{\boldsymbol{\prime\prime}}$	$\frac{\hat{\boldsymbol{\prime\prime}}^T\hat{\boldsymbol{\prime\prime}}}{p}$	σ^2
Total	n	$\mathbf{y}^T\mathbf{y}$		

Table 4: Analysis Of Variance For The General Linear Model

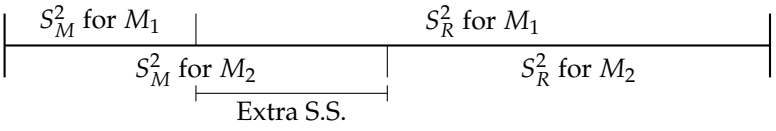

Figure 3: Diagrammatic representation of the extra sum of squares.

The information deduced above is summarised in Table 18.2.3.

The Extra Sum Of Squares Principle

It is all very well fitting a model, but what if it is inadequate? Or what if it is too elaborate, and a simpler model would do? Suppose that a model M_1 is fitted, and that it is a special case of a more general model M_2. Clearly fitting M_2 will cause the residual sum of squares to decrease, and the amount by which it does so is the 'extra' sum of squares. This is illustrated in Figure 18.2.3.

Theoretically, the extra sum of squares can be represented in an Analysis Of Variance table. First of all, since "$M_1 \subset M_2$",

Model 30 *A model can be augmented by the introduction of new parameters.*

$$M_1: \mathbf{y} = X\grave{} + \boldsymbol{\prime\prime}$$
$$M_2: \mathbf{y} = X\grave{} + X^*\grave{}^* + \boldsymbol{\prime\prime}$$

In matrix form, M_2 may be expressed as

$$\mathbf{y} = \begin{pmatrix} X & X^* \end{pmatrix} \begin{pmatrix} \grave{} \\ \grave{}^* \end{pmatrix} + \boldsymbol{\prime\prime}$$

By considering each of the models separately, the sum of squares can now be decomposed as one of

$$\mathbf{y}^T\mathbf{y} = \hat{\mathbf{y}}_1^T\hat{\mathbf{y}}_1 + \left(\hat{\boldsymbol{\prime\prime}}_1^T\hat{\boldsymbol{\prime\prime}}_1 - \hat{\boldsymbol{\prime\prime}}_2^T\hat{\boldsymbol{\prime\prime}}_2\right) + \hat{\boldsymbol{\prime\prime}}_2^T\hat{\boldsymbol{\prime\prime}}_2 \tag{31}$$

$$\text{and} \quad \mathbf{y}^T\mathbf{y} = \hat{\mathbf{y}}_1^T\hat{\mathbf{y}}_1 + \left(\hat{\mathbf{y}}_2^T\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1^T\hat{\mathbf{y}}_1\right) + \hat{\boldsymbol{\prime\prime}}_2^T\hat{\boldsymbol{\prime\prime}}_2 \tag{32}$$

In both cases the quantity in the brackets is the extra sum of squares. Table 18.2.3 shows how the resulting Analysis Of Variance is modified: the top part of the table shown the breakdown of M_2.

Notice that $S_{M_1}^2 = S_{M_2}^2 + ESS$. The appropriate $\mathcal{F}$ test ratio is therefore $\frac{ESS}{S_{M_2}^2}$. It is therefore usual to abbreviate the Analysis Of Variance table, as is shown in Table 18.2.3

Source	D.F.	Sum Of Squares
M_1	p_1	$\hat{\mathbf{y}}_1^T\hat{\mathbf{y}}_1$
M_2 adjusted for M_1, the E.S.S.	$p_2 - p_1$	$\hat{\mathbf{y}}_2^T\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1^T\hat{\mathbf{y}}_1$
M_2	p_2	$\hat{\mathbf{y}}_2^T\hat{\mathbf{y}}_2$
Residual (from M_2)	$n - p_2$	$\hat{\boldsymbol{\varepsilon}}_2^T\hat{\boldsymbol{\varepsilon}}_2$
Total	n	$\mathbf{y}^T\mathbf{y}$

Table 5: Analysis Of Variance showing the extra sum of squares.
This analysis of variance uses equation (31); the sum of the M_1 sum of squares and the extra sum of squares is the M_2 sum of squares.

Source	D.F.	Sum Of Squares
Extra Sum Of Squares	$p_2 - p_1$	$\hat{\mathbf{y}}_2^T\hat{\mathbf{y}}_2 - \hat{\mathbf{y}}_1^T\hat{\mathbf{y}}_1$
M_2 Residual	$n - p_2$	$\hat{\boldsymbol{\varepsilon}}_2^T\hat{\boldsymbol{\varepsilon}}_2$
Adjusted Total = M_1 Residual	$n - p_1$	$\mathbf{y}^T\mathbf{y} - \hat{\mathbf{y}}_1^T\hat{\mathbf{y}}_1$

Table 6: Adjusted Analysis Of Variance showing the extra sum of squares.
This analysis of variance uses equation (32). The significance of the extra sum of squares can be tested using the $\mathcal{F}$ statistic arrising from the main part of the table.

Application To Linear Regression

For the general model take

$$\mathbf{y} = X\begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \varepsilon$$

and the specific case to be $\beta = 0$.

Clearly, under M_1 $\hat{alpha} = \overline{y}$ and hence the model sum of squares is

$$\begin{aligned}\sum_{i=1}^{n}\hat{y}_i^2 &= \sum_{i=1}^{n}\overline{y} \quad \text{because } \hat{y} = \overline{y}\ \forall i\\ &= n\overline{y}^2\end{aligned}$$

Under M_2, $\hat{\beta} = \frac{C_{xy}}{C_{xx}}$ and $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$. Hence,

$$\begin{aligned}\sum_{i=1}^{n}\hat{y}_i^2 &= \sum_{i=1}^{n}\left(\overline{y} - \hat{\beta}\overline{x} + \hat{\beta}x_i\right)^2\\ &= \sum_{i=1}^{n}\overline{y}^2 + 2\overline{y}\hat{\beta}(\overline{x} - x_i) + \hat{\beta}^2(\overline{x} - x_i)^2\\ &= n\overline{y}^2 + \hat{\beta}^2 C_{xx}\end{aligned}$$

Hence the complete analysis of variance table can be deduced and is as shown in Table 18.2.3

Note that the extra sum of squares effectively *is* the regression sum of squares.

Application To The k Sample Problem

Consider the models

$$M_1 : y_{ij} = \mu + \varepsilon_{ij} \qquad M_2 : y_{ij} = \mu + \theta_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}$$

Source	D.F.	Sum Of Squares
M_1	1	$n\overline{y}^2$
Extra Sum Of Squares	1	$\hat{\beta}C_{xx}$
M_2	2	$n\overline{y}^2 + \hat{\beta}C_{xx}$
Residual	$n-2$	Find by subtraction
Total	n	$\sum_{i=1}^{n} y_i^2$

Table 7: Complete Analysis Of Variance For Linear Regression Models

Clearly for M_1, $\hat{\mu} = \overline{y}$. For M_2 however, $\hat{\mu}_i = \overline{y}_i$.

Now, for M_1

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \hat{y}_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \overline{y} = N\overline{y}^2$$

and for M_2,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \hat{y}_{ij}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \overline{y}_i^2 = \sum_{i=1}^{k} n_i\overline{y}_i^2$$

The extra sum of squares is therefore

$$\sum_{i=1}^{k} n_i\overline{y}_i^2 - N\overline{y}^2 = \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2$$

which is the between groups sum of squares. The complete analysis of variance is given in Table 18.2.3.

Source	D.F.	Sum Of Squares
M_1	1	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \overline{y}$
Extra Sum Of Squares	$k-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (\overline{y} - \overline{y}_i)^2$
M_2	k	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \overline{y}_i$
Residual	$N-k$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$
Total	N	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2$

Table 8: Complete analysis of variance for the k sample problem.

The abbreviated analysis of variance table produces the usual k sample problem analysis of variance.

k Group Regression

A hybrid between linear regression and the k sample problem is k group regression. Each x_i has associated with it many rather than just one y value, so

$$M_1 : y_{ij} = \alpha + \beta x_i + \varepsilon_{ij} \qquad M_2 : y_{ij} = \mu_i + \varepsilon_{ij}$$

M_2 has k parameters whereas M_1 has only 2, this is because M_1 presumes that there is a (linear) relationship between the μ_is and so eliminates these parameters. Hence the required setup exists for an application of the extra sum of squares principle. If the extra sum of squares is not significant, then M_1 is sufficient to describe accurately the data, so the relationship is linear.

Source	D.F.	Sum Of Squares
x_1	1	Sequential sum of squares 1, a
x_2	1	Sequential sum of squares 2, b
x_3	1	Sequential sum of squares 3, c
x_4	1	Sequential sum of squares 4, d
Regression	3	$a+b+c+d$
Error		
Total		

Table 9: Decomposition of regression sum of squares into sequential sum of squares.

Multiple Regression

In linear regression the data y_i are accounted for by a single explanatory variable x. This is now extended to many explanatory variables $x_1, x_2, \ldots, x_k$ so that

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

This model is a local approximation to the hyperplane $y = f(\mathbf{x}) + "$, and can be formulated as a general linear model with design matrix

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}$$

By letting say $x_r = x_1^r$ etc. it is possible to test for polynomial relationships.

Having fitted the model, the $\mathcal{F}$ test will test the hypothesis

$$H_0 \colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Statistical computer software may often quote "r^2", which is the percentage of the total sum of squares that is accounted for by the model.

The more variables in the model, the more of the variability can be explained, however, to include many variables is clearly undesirable. It is required to find out the minimum number of explanatory variables required in order to effectively describe the data. When the estimates of the βs are calculated and their variances found, a t test can be used to test the hypothesis $\beta_j = 0$ regardless of the other βs. This is a good way to see if some of the βs can be dropped, but there is a problem with this: Only one β can be removed at once as its partition of the sum of squares will need to be re-distributed and so once not-significant βs may well become significant.

It would be possible to try eliminating a particular β and then re-test, but there is a better way. A simple model with only one explanatory variable can be used initially, then more explanatory variables added in one by one. Each time an explanatory variable is added in there will be an extra sum of squares—these are the sequential sums of squares. In this case the order in which the explanatory variables are introduced will make a difference to the sequential sums of squares.

Further $\mathcal{F}$ tests can be performed by considering the sum of some of the sums of squares. For example take a situation as shown in Table 18.2.3, then the influence of x_3 and x_4 can be tested using an $\mathcal{F}$ test on $\frac{c+d}{\text{Error S.S.}}$.

(18.3) Dependent Random Variables

(18.3.1) Covariance & Correlation

Definition & Properties

Definition 33 *Let X and Y be random variables with $\mathbb{E}\,X = \mu_x$, $\mathbb{E}\,Y = \mu_y$, $\operatorname{var} X = \sigma_x^2$, and $\operatorname{var} Y = \sigma_y^2$. The covariance of X and Y is defined by*

$$\operatorname{cov}(X,Y) = \mathbb{E}\left((X-\mu_x)(Y-\mu_y)\right)$$

The correlation of X and Y is defined by

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_x\sigma_y}$$

Straight from the definition it is clear that

1. $\operatorname{cov}(X,Y) = \mathbb{E}(XY) - \mu_x\mu_y$, which can be shown by multiplying out the definition.
2. $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$.
3. $\operatorname{cov}(X,c) = 0$ where c is a constant.
4. $\operatorname{cov}(X,X) = \operatorname{var} X$.

A linearity property follows from the linearity of the expectation operator; bearing in mind that if $\mathbb{E}\,X = \mu_x$ then $\mathbb{E}(aX) = a\mu_x$,

$$\begin{aligned}\operatorname{cov}(aX+bY,Z) &= \mathbb{E}(aX+bY-a\mu_u-b\mu_y)(Z-\mu_z)\\ &= \mathbb{E}(aX-a\mu_x)(Z-\mu_z) + \mathbb{E}(bY-b\mu_y)(Z-\mu_z)\\ &= a\operatorname{cov}(X,Z) + b\operatorname{cov}(Y,Z)\end{aligned}$$

Using this property together with $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$ it can be shown that covariance is a bilinear form, i.e.

$$\operatorname{cov}\left(\sum_i a_iX_i, \sum_j b_jY_j\right) = \sum_i\sum_j a_ib_j\operatorname{cov}\left(X_i,Y_j\right)$$

If X and Y are independent random variables, then $\mathbb{E}(XY) = (\mathbb{E}\,X)(\mathbb{E}\,Y)$, from which it follows that $\operatorname{cov}(X,Y) = 0$. However, the converse is not true. The merit of covariance, therefore, is the ability to deal with dependent random variables. In particular it should be noted that

$$\begin{aligned}\operatorname{var}(aX+bY) &= \mathbb{E}(aX+bY-a\mu_x-b\mu_y)^2\\ &= \mathbb{E}(aX-a\mu_x)^2 + \mathbb{E}(bY-b\mu_y)^2 + 2\,\mathbb{E}(aX-\mu_x)(bY-\mu_y)\\ &= a^2\operatorname{var}X + b^2\operatorname{var}Y + 2ab\operatorname{cov}(X,Y)\end{aligned}$$

$$\text{and so } \operatorname{var}\left(\sum_i a_iX_i\right) = \sum_i a_i^2\operatorname{var}X_i + 2\sum_{i<j} a_ia_jX_iX_j$$

Theorem 34 *For any random variables X and Y with non-zero variance,*

$$-1 \leqslant \rho(X,Y) \leqslant 1$$

with equality if and only if $Y = cX + d$ for constants c and d, and when this is the case, $\rho = \operatorname{sgn} c$.

Proof. Let α and β be constants, then

$$\begin{aligned} 0 \leqslant \operatorname{var}(\alpha X + \beta Y) &= \alpha^2 \operatorname{var} X + \beta^2 \operatorname{var} Y + 2\alpha\beta \operatorname{cov}(X, Y) \\ &= \alpha^2\sigma_x^2 + \beta^2\sigma_y^2 + 2\alpha\beta\sigma_x\sigma_y\rho \end{aligned}$$

Each of the inequalities is shown separately. Since the relationship must hold for all α and β, they can be chosen at will, hence

- Put $\alpha = \frac{1}{\sigma_x}$ and $\beta = \frac{1}{\sigma_y}$, which gives

$$0 \leqslant 1 + 1 + 2\rho \quad \Rightarrow \quad \rho \geqslant -1$$

- Put $\alpha = \frac{1}{\sigma_x}$ and $\beta = \frac{-1}{\sigma_y}$, which gives

$$0 \leqslant 1 + 1 - 2\rho \quad \Rightarrow \quad \rho \leqslant 1$$

Now, if $\rho = \pm 1$,

$$\rho = \pm 1 \quad \Leftrightarrow \quad \operatorname{var}(\alpha X + \beta Y) = \alpha^2\sigma_x^2 + \beta^2\sigma_y^2 \pm 2\alpha\beta\sigma_x^2\sigma_y^2$$

now put $\alpha = \frac{1}{\sigma_x}$ and $\beta = \frac{\mp 1}{\sigma_y}$ to give

$$\begin{aligned} \Rightarrow \quad & \operatorname{var}\left(\frac{X}{\sigma_x} \mp \frac{Y}{\sigma_y}\right) = 1 + 1 - 2 = 0 \\ \Rightarrow \quad & \frac{X}{\sigma_x} \mp \frac{Y}{\sigma_y} = d \quad \text{for some constant } d \\ \Rightarrow \quad & Y = \pm\frac{\sigma_y}{\sigma_x} X + d_1 \end{aligned}$$

Hence when $\rho = \pm 1$, $Y = cX + d$ with $\rho = \operatorname{sgn} c$. What now remains to be shown is the reverse implication. Say $Y = \pm cX + d$, so

$$\operatorname{cov}(X, Y) = \operatorname{cov}(X, \pm cX + d) = \pm c \operatorname{cov}(X, X) + \operatorname{cov}(X, d) = \pm c \operatorname{var} X = c\sigma_x^2$$

But since $Y = \pm cX + d$, $\sigma_y^2 = c^2\sigma_x^2$ so,

$$\rho = \frac{\operatorname{cov}(X, Y)}{\sigma_x\sigma_y} = \frac{\pm c\sigma_x^2}{|c|\sigma_x^2} = \pm 1$$

Hence the result. □

Covariance Matrices

The study of many random variables at once is much simplified by exploiting matrix algebra—this has already been seen with the General Linear model. Suppose that the (not necessarily independent) random variables $X_1, X_2, \ldots X_n$ are the elements of an $n \times 1$ vector $\mathbf{X}$.

- The expected value of $\mathbf{X}$ is $\bar{} = \mathbb{E}\,\mathbf{X}$ where $\mu_i = \mathbb{E}\,X_i$.

- The covariance matrix of X is the $n \times n$ matrix V with $V_{ij} = \operatorname{cov}\left(X_i, X_j\right)$. So

$$V = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \dots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \dots \sigma_n^2 & \end{pmatrix}$$

  Observe that V is symmetric, and if the X_is are independent, then V will be diagonal.

Theorem 35 *Suppose the random vector variable $\mathbf{X}$ has mean $\bar{}$ and covariance matrix V. Then*

1. *If $\mathbf{a}$ is a constant $n \times 1$ vector and $Y = \mathbf{a}^T\mathbf{X}$ then $\mathbb{E}\, Y = \mathbf{a}^T\bar{}$ and $\operatorname{var} Y = \mathbf{a}^T V\mathbf{a}$.*
2. *If A is a constant $m \times n$ matrix then where $\mathbf{W} = A\mathbf{X}$, $\mathbb{E}\,\mathbf{W} = A\bar{}$ and the covariance matrix of $\mathbf{W}$ is AVA^T.*

Proof. 1. For the mean,

$$\begin{aligned} \mathbb{E}\, Y = \mathbb{E}\,(\mathbf{a}^T\mathbf{X}) &= \mathbb{E} \sum_{i=1}^{n} a_i X_i \\ &= \sum_{i=1}^{n} a_i\, \mathbb{E}\, X_i \\ &= \sum_{i=1}^{n} a_i \mu_i \\ &= \mathbf{a}^T\bar{} \end{aligned}$$

For the variance,

$$\begin{aligned} \operatorname{var} Y = \operatorname{cov}(Y, Y) &= \operatorname{cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} a_j X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{cov}\left(X_i, X_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i V_{ij} a_j \\ &= \mathbf{a}^T V\mathbf{a} \end{aligned}$$

2. For the mean,

$$\mathbb{E}\, W_i = \mathbb{E} \sum_{j=1}^{n} A_{ij} W_j = \sum_{j=1}^{n} A_{ij}\, \mathbb{E}\, W_j = A\bar{}$$

For the covariance matrix,

$$\begin{aligned} \operatorname{cov}\mathbf{W} &= \mathbb{E}\left((\mathbf{W} - A\bar{})(\mathbf{W} - A\bar{})^T\right) \\ &= \mathbb{E}\left((A\mathbf{X} - A\bar{})(A\mathbf{X} - A\bar{})^T\right) \\ &= A\,\mathbb{E}\left((\mathbf{X} - \bar{})(\mathbf{X} - \bar{})^T\right) A^T \\ &= AVA^T \end{aligned}$$

□

Application To The General Linear Model

The General Linear model is of the form $\mathbf{y} = X\grave{} + ''$, and clearly the covariance matrix of $''$ is $\mathbb{E}\,(''''^T) = \sigma^2 I$.

The parameter ` is estimated by ˆ = $\left(X^TX\right)^{1-} X^T\mathbf{y}$, the covariance matrix of which is found using the second part of Theorem 35. Firstly,

$$\begin{aligned}\operatorname{cov}\mathbf{y} &= \mathbb{E}\,(\mathbf{y}-\bar{}_y)(\mathbf{y}-\bar{}_y)^T \quad \text{but } \bar{}_y = \mathbb{E}\,\mathbf{y} = X` \\ &= \mathbb{E}\,\prime\prime\prime^T \\ &= \sigma^2 I\end{aligned}$$

Now using Theorem 35,

$$\begin{aligned}\operatorname{cov}\hat{} &= \left(\left(X^TX\right)^{-1}X^T\right)\sigma^2 I\left(\left(X^TX\right)^{-1}X^T\right)^T \\ &= \sigma^2\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1} \\ &= \sigma^2\left(X^TX\right)^{-1}\end{aligned}$$

This confirms the result on page 17.

Similarly the covariance matrix of $\hat{\prime\prime}$ is

$$\sigma^2(I-H)I(I-H)^T = \sigma^2(I-H)$$

from which it is evident that $\operatorname{var}\varepsilon_i = \sigma^2(1-h_{ii})$.

Application To Linear Regression

The linear model for regression is

$$\mathbf{y} = \begin{pmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \prime\prime$$

from which,

$$V = \operatorname{cov}\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = \frac{\sigma^2}{nC_{xx}}\begin{pmatrix}\sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i\\ -\sum_{i=1}^n x_i & n\end{pmatrix}$$

From this it is evident that

$$\operatorname{var}\hat{\alpha} = \frac{\sigma^2\sum_{i=1}^n x_i^2}{nC_{xx}} \quad \operatorname{var}\hat{\beta} = \frac{\sigma^2}{C_{xx}} \quad \operatorname{cov}\left(\hat{\alpha},\hat{\beta}\right) = \frac{-\sigma^2\overline{x}}{C_{xx}}$$

This means that if $\overline{x}$ is located at the origin then there is no covariance between $\hat{\alpha}$ and $\hat{\beta}$. This is because the slope of the line effects the position of the intercept, unless the intercept is at the origin.