MATH 2016 (13177) STATISTICAL MODELLING

In this chapter a number of topics are revised to make sure that we all have a basic understanding. I am going to assume that you are thoroughly familiar with the basic matrix algebra and the use of summation notation: $\sum_{i=1}^{n}\ldots$ If unsure try it out on a small arithmetical example that you can easily verify. For example to show that $\sum_i ax_i = a\sum_i x_i$, verify for a = 2 and x_1 = 1 and x_1 = 3: $(2\times 1)+(2\times 3)=2\times(1+3)$.

I. Statistical inference

I.A Expected values and variances

As you may know, statistical inference is about drawing conclusions about one or more populations based on samples from the populations. Generally, we compute numeric quantities from the sample (means, variances, proportions) and these are called **statistics** or **estimates**. They are used as estimates of particular population quantities, these being called **parameters**. It is very important to be clear about whether parameters (populations) or statistics (samples) are being referred to. Thus when one is talking about a mean, is it the population or sample mean? To aid in making the distinction, the convention adopted is to use Greek letters as symbols for parameters and ordinary Roman letters as symbols for statistics.

Fundamental to this course are the concepts of population expected value and variance. The expected value is the mean of the variable Y in a population — it is a population parameter.

Definition I.1: The **expected value**, $\psi_Y = E[Y]$, of a continuous random variable Y whose population distribution is described by $f(y)$ is given by

$$E[Y] = \int_{-\infty}^{\infty} yf(y)dy$$

■

That is, $\psi_Y = E[Y]$ is the mean in a population whose distribution is described by $f(y)$.

Theorem I.1: Let Y be a continuous random variable with probability distribution function $f(y)$. The expected value of a function $u(Y)$ of the random variable is

$$E\left[u(Y)\right] = \int_{-\infty}^{\infty} u(y) f(y)\, dy .$$

Proof: not given ■

Note that any function of a random variable is itself a random variable. This theorem tells us how to get the expected value of the function. It will now be used to find $E\left[a \times v(Y) + b\right]$ by setting $u(Y) = a \times v(Y) + b$.

Theorem I.2: $E\left[a \times v(Y) + b\right] = aE\left[v(Y)\right] + b$

Proof:

For a continuous random variable, we have from theorem I.1

$$\begin{aligned} E\left[a \times v(Y) + b\right] &= \int_{-\infty}^{\infty} \left\{a \times v(y) + b\right\} f(y)\, dy \\ &= \int_{-\infty}^{\infty} \left\{a \times v(y)\right\} f(y)\, dy + \int_{-\infty}^{\infty} bf(y)\, dy \\ &= a\int_{-\infty}^{\infty} v(y) f(y)\, dy + b\int_{-\infty}^{\infty} f(y)\, dy \\ &= aE\left[v(Y)\right] + b \end{aligned}$$

■

In particular, $E\left[aY + b\right] = aE\left[Y\right] + b$.

Definition I.2: The **variance**, $\text{var}\left[Y\right] = \sigma_Y^2$, of any random variable Y is defined to be

$$\text{var}\left[Y\right] = \sigma_Y^2 = E\left[\left(Y - \psi_Y\right)^2\right] = E\left[\left(Y - E\left[Y\right]\right)^2\right]$$

■

That is variance is the mean in the population of the squares of the deviations of the observed values from the population mean — it measures how far on average observations are from the mean in the population. It is also a population parameter.

Theorem I.3: The variance, $\text{var}\left[Y\right] = \sigma_Y^2$, of a continuous random variable Y whose population distribution is described by $f(y)$ is given by

$$\text{var}\left[Y\right] = \sigma_Y^2 = \int_{-\infty}^{\infty} \left(y - \psi_Y\right)^2 f(y)\, dy .$$

Proof: This is a straight forward application of theorem I.1 where $u(Y) = (Y - \psi_Y)^2$. ■

In this course, as in much of Statistics, it is common to assume that a continuous variable is normally distributed. The distribution function for such a variable involves the parameters ψ_Y and σ_Y as follows:

$$f(y) = \frac{1}{\sqrt{(2\pi\sigma_Y^2)}} \exp\left\{\frac{-(y-\psi_Y)^2}{2\sigma_Y^2}\right\}$$

Clearly, to determine the particular normal distribution that is to apply in our example, the values of ψ and σ^2 (or σ) must be specified. (Note we drop the subscript *Y*, taking it as implicit.) We require an estimator of ψ (and perhaps σ). Here we concentrate on the estimator of ψ. Suppose we have a sample $y_1, y_2, \ldots, y_n$. Note the lower case *y* for observed values as opposed to *Y* for the random variable. The obvious estimator of ψ is the sample mean $\bar{Y} = \sum_{i=1}^{n} Y_i / n$. Note we call the formula that tells us how to estimate a parameter an **estimator** and it is a function of random variables, *Y*s. The value obtained by substituting the sample values into the formula is called the **estimate** and it is a function of observed values, *y*s.

It is common practice to denote the estimator as the parameter with a caret over it. For example, $\hat{\psi} = \bar{Y}$ means that the estimator of ψ is $\bar{Y}$; in this notation $\hat{\psi}$ also stands for the estimate so that $\hat{\psi} = \bar{y}$ means that an estimate of ψ is $\bar{y}$.

I.B The linear regression model

In this section we consider models of the general form:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p + \varepsilon$$

where *Y* is a continuous random variable and x_is are quantitative variables that are called the explanatory variables.

This model is called a linear model in that it is linear in the θ_is. Consider the following models.

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \varepsilon$$
$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2 + \varepsilon$$
$$Y = \theta_0 + \theta_1 \ln(x_1) + \varepsilon$$
$$Y = e^{\theta_0 x_1} + \varepsilon$$
$$Y = 1/\left(1 + e^{-\theta_0 - \theta_1 x_1 + \varepsilon}\right)$$

All but the last two are linear in the θ_is.

We would conduct a study in which n ($\geq p$+1) observations are taken of the response variable Y and the explanatory variables, x_is. This leads to the following system of equations that model the observed responses.

$$\begin{aligned}
Y_1 &= \theta_0 + \theta_1 x_{11} + \theta_2 x_{12} + \ldots + \theta_p x_{1p} + \varepsilon_1 \\
Y_2 &= \theta_0 + \theta_1 x_{21} + \theta_2 x_{22} + \ldots + \theta_p x_{2p} + \varepsilon_2 \\
&\vdots \\
Y_i &= \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} + \varepsilon_i \\
&\vdots \\
Y_n &= \theta_0 + \theta_1 x_{n1} + \theta_2 x_{n2} + \ldots + \theta_p x_{np} + \varepsilon_n
\end{aligned}$$

What does the model tell us about our data? Well, we have a response variable, Y, whose values are related to several explanatory variables, x_is. Note the use of lower case x for the explanatory variables to signify that they are not random variables — their values are considered to be known without error. However, we do not always get the same value of the response variable when we observe the same combination of values of the explanatory variables — these differences are accounted for by the ε_is, the **random errors**, in the above model.

It is also usual to make further assumptions about ε_is: $E\left[\varepsilon_i\right] = 0$, $\text{var}\left[\varepsilon_i\right] = \sigma^2$ and $\text{cov}\left[\varepsilon_i, \varepsilon_j\right] = 0,\ i \neq j$. The assumptions about the ε_is mean that on average the errors cancel out so that we get the population value of the response, that the variability of the errors is independent of the values of any of the variables and that the error in one observation is unrelated to that of any other observation.

The last assumption involves a third quantity involving expectations: covariance.

Definition I.3: The covariance of two random variables, X and Y, is defined to be

$$\text{cov}\left[X,Y\right] = E\left[\left(X - E\left[X\right]\right)\left(Y - E\left[Y\right]\right)\right]$$

■

The covariance measure the extent to which the two random variables values move together. In fact, the linear correlation coefficient can be calculated from it as follows:

$$\text{corr}\left[X,Y\right] = \frac{\text{cov}\left[X,Y\right]}{\sqrt{\text{var}\left[X\right]\text{var}\left[Y\right]}}$$

That is, the correlation coefficient is just the covariance adjusted for the variance of X and Y.

We can express the system of equations in terms of matrices. However, there is a slight notational glitch in that it is usual to indicate matrices by bolded upper case letters and vectors by bolded lower case. We will follow this notation except that vectors of random variables will also be in upper case. Before obtaining expressions

for the system of equations we need definitions of the expectation and variance of a random vector.

Definition I.4: Let **Y** be a vector of *n* jointly-distributed random variables with $E[Y_i] = \psi_i$, $\text{var}[Y_i] = \sigma_i^2$ and $\text{cov}[Y_i, Y_j] = \sigma_{ij} \left(= \sigma_{ji}\right)$. Then, the **random vector** is

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

The **expectation vector**, $\boldsymbol{\psi}$, giving the expectation of **Y** is

$$E[\mathbf{Y}] = \begin{bmatrix} E[Y_1] \\ E[Y_2] \\ \vdots \\ E[Y_n] \end{bmatrix} = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{bmatrix} = \boldsymbol{\psi}$$

The **variance matrix**, **V**, giving the variance of **Y** is

$$\mathbf{V} = E\left[\left(\mathbf{Y} - E[\mathbf{Y}]\right)\left(\mathbf{Y} - E[\mathbf{Y}]\right)'\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1j} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2i} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{1j} & \sigma_{2i} & \cdots & \sigma_j^2 & \cdots & \sigma_{jn} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{in} & \cdots & \sigma_n^2 \end{bmatrix}$$

■

Before returning to our system of equations we note that in the last expression we have the transpose operator. This operator is going to occur so frequently that at this point we make an aside to examine its properties, which de via the following lemma.

Lemma I.1: The transpose of a matrix displays the following properties:

1. The transpose of a transpose yield the original result — $\left(\mathbf{A}'\right)' = \mathbf{A}$.
2. The transpose of a column vector is a row vector and vice versa so that we always write the column vector as untransposed and the row vector as transposed — **a** is a column vector and **a**′ is the corresponding row vector.
3. The transpose of a linear combination of matrices is the same linear combination of the transposes — $\left(c\mathbf{A} + d\mathbf{B}\right)' = c\mathbf{A}' + d\mathbf{B}'$.
4. The transpose of a product is the product of the transposes, but with the order of the matrices reversed — $\left(\mathbf{AB}\right)' = \mathbf{B}'\mathbf{A}'$.
5. Transposing a symmetric matrix leaves the matrix unchanged — $\mathbf{A}' = \mathbf{A}$ provided **A** is symmetric.
6. The product of two different symmetric matrices is not symmetric — $\left(\mathbf{AB}\right)' = \mathbf{BA}$ for **A** and **B** symmetric.
7. The product of any matrix with its transpose is symmetric — $\left(\mathbf{A}'\mathbf{A}\right)' = \mathbf{A}'\left(\mathbf{A}'\right)' = \mathbf{A}'\mathbf{A}$ and $\left(\mathbf{AA}'\right)' = \left(\mathbf{A}'\right)'\mathbf{A}' = \mathbf{AA}'$.

8. A column vector premultiplied by a row vector is a scalar and so is symmetric — $\mathbf{a}'\mathbf{b} = \sum_{i=1}^{n} a_i b_j$ and $\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = \left(\mathbf{a}'\mathbf{b}\right)'$.
9. A column vector premultiplied by its transpose is the sum of squares of its elements, also a scalar — $\mathbf{a}'\mathbf{a} = \sum_{i=1}^{n} a_i^2$.
10. A column vector of order n post multiplied by its transpose is a symmetric matrix of order $n \times n$ — from property 7 we have $\left(\mathbf{a}\mathbf{a}'\right)' = \mathbf{a}\mathbf{a}'$.

The most important of these properties at this stage are 1, 2, 4, 9 and 10.

In particular, property 10 applies to **V** in definition I.4 and tells us that **V** is an $n \times n$ symmetric matrix.

Now returning to our system of equations, to express them in matrix terms let,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{i1} & x_{12} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \ \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}, \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The system of equations can be written using this notation as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

with $E\left[\boldsymbol{\varepsilon}\right] = \mathbf{0}$ and $\text{var}\left[\boldsymbol{\varepsilon}\right] = \mathbf{V}_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbf{I}_n$ where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

An alternative way to express this model is in terms of $E\left[\mathbf{Y}\right]$ and $\text{var}\left[\mathbf{Y}\right]$. These alternative expressions can be obtained by substituting the model into $E\left[Y_i\right]$, $\text{var}\left[Y_i\right]$ and $\text{cov}\left[Y_i, Y_j\right]$. Thus,

$$\begin{aligned} E\left[Y_i\right] &= E\left[\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} + \varepsilon_i\right] \\ &= \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} + E\left[\varepsilon_i\right], \\ &= \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} \end{aligned}$$

$$\begin{aligned} \text{var}\left[Y_i\right] &= E\left[\left(Y_i - E\left[Y_i\right]\right)^2\right] \\ &= E\left[\left(\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_p x_{ip} + \varepsilon_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - \ldots - \theta_p x_{ip}\right)^2\right] \\ &= E\left[\varepsilon_i^2\right] \\ &= \sigma^2 \qquad \left(\text{since } \text{var}\left[\varepsilon_i\right] = E\left[\left(\varepsilon_i - E\left[\varepsilon_i\right]\right)^2\right] = E\left[\varepsilon_i^2\right]\right) \end{aligned}$$

and

$$\begin{aligned}\text{cov}\left[Y_i, Y_j\right] &= E\left[\left(Y_i - E\left[Y_i\right]\right)\left(Y_j - E\left[Y_j\right]\right)\right] \\ &= E\left[\varepsilon_i \varepsilon_j\right] \\ &= \text{cov}\left[\varepsilon_i, \varepsilon_j\right] \\ &= 0\end{aligned}$$

In matrix terms, the alternative expression for the model is:

$$E\left[\mathbf{Y}\right] = \mathbf{X\theta} \text{ and } \text{var}\left[\mathbf{Y}\right] = \mathbf{V_Y} = \sigma^2 \mathbf{I}_n .$$

That is, $\mathbf{V_\varepsilon}$ is also the variance matrix for **Y**.

Example I.1 House price

Suppose it is thought that the price obtained for a house depends primarily the age and livable area. We observe 5 randomly selected houses on the market and obtain the following data:

Price \$'000 (y)	Age years (x_1)	Area '000 feet^2 (x_2)
50	1	1
40	5	1
52	5	2
47	10	2
65	20	3

In this example, n = 5 and p = 2. The model that we propose for this data is as follows:

$$Y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \varepsilon_i$$

with $E\left[\varepsilon_i\right] = 0$, $\text{var}\left[\varepsilon_i\right] = \sigma^2$ and $\text{cov}\left[\varepsilon_i, \varepsilon_j\right] = 0,\ i \neq j$,

or, equivalently,

$$E\left[Y_i\right] = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2}$$

with $\text{var}\left[Y_i\right] = \sigma^2$ and $\text{cov}\left[Y_i, Y_j\right] = 0,\ i \neq j$,

In matrix terms, the model is:

$$\mathbf{Y} = \mathbf{X\theta} + \boldsymbol{\varepsilon} \text{ with } E\left[\boldsymbol{\varepsilon}\right] = \mathbf{0} \text{ and } \text{var}\left[\boldsymbol{\varepsilon}\right] = \mathbf{V} = \sigma^2 \mathbf{I}_n ,$$

or, equivalently,

$$E\left[\mathbf{Y}\right] = \mathbf{X\theta} \text{ and } \text{var}\left[\mathbf{Y}\right] = \mathbf{V} = \sigma^2 \mathbf{I}_n$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 10 & 2 \\ 1 & 20 & 3 \end{bmatrix}, \ \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}, \ \mathbf{V} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

We also have the vector, **y**, of observed values of **Y**:

$$\mathbf{y} = \begin{bmatrix} 50 \\ 40 \\ 52 \\ 47 \\ 65 \end{bmatrix}$$

■

Example I.2 Voter turnout

In this example a political scientist attempted to investigate, following an election, the relationship between campaign expenditures on televised advertisements and subsequent voter turnout. The aim is to be able to predict voter turnout from advertising expenditure. That is, voter turnout is the response or dependent variable and is denoted by *Y*; advertising expenditure is the explanatory or independent variable and is denoted by *x*. The following table presents the percent of total campaign expenditures relegated to televised advertisements and the percent of registered voter turnout for a sample of 20 electorates.

Voter Turnout	% Advert Expenditure	Voter Turnout	% Advert Expenditure
35.4	28.5	40.8	31.3
58.2	48.3	61.9	50.1
46.1	40.2	36.5	31.3
45.5	34.8	32.7	24.8
64.8	50.1	53.8	42.2
52.0	44.0	24.6	23.0
37.9	27.2	31.2	30.1
48.2	37.8	42.6	36.5
41.8	27.2	49.6	40.2
54.0	46.1	56.6	46.1

This example is one that involves simple linear regression as there is only one explanatory variable. In this case, we drop the subscript for the independent variable so that our proposed model is:

$$E\left[Y_i\right] = \theta_0 + \theta_1 x_i$$

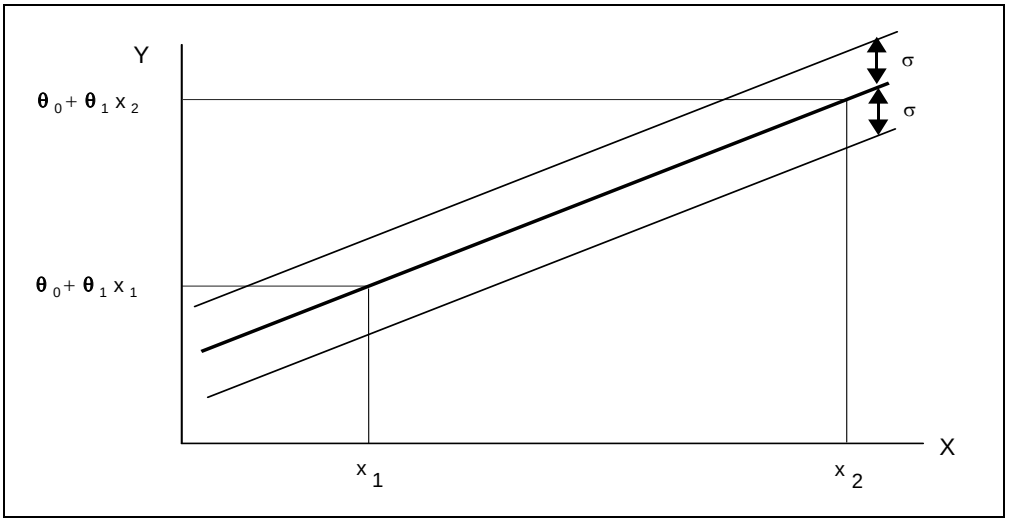
with $\text{var}\left[Y_i\right] = \sigma^2$ and $\text{cov}\left[Y_i, Y_j\right] = 0, \ i \neq j$.

This model is to be used to represent the way in which it is suggested the data behaves. So how should it behave for this model?

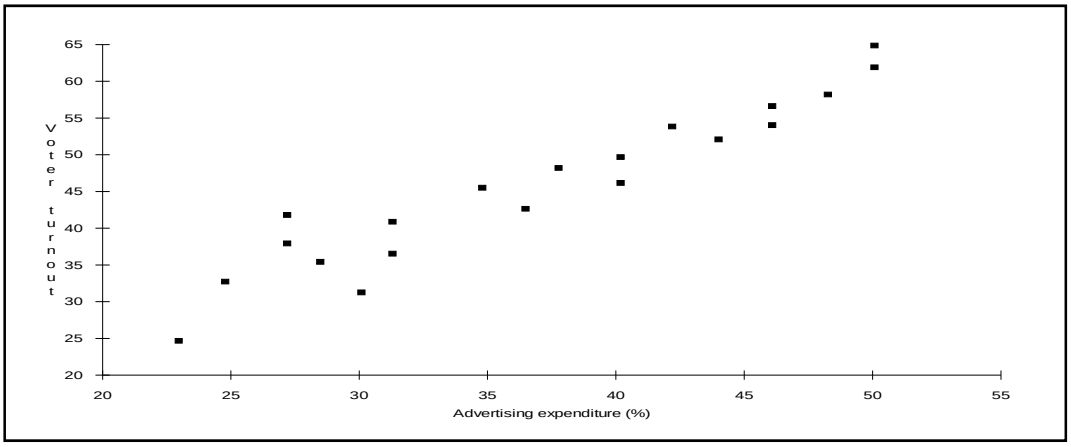
$E[Y_i]$ means the population average value or population mean response. The equation given above for $E[Y_i]$ implies that the average value of *Y*, for individuals for which the value of *x* is x_i, is $\theta_0 + \theta_1 x_i$. Thus, if the value of *x* for a particular observation that the investigator takes is x_i, then it would be expected that the observed value of *Y* would be $\theta_0 + \theta_1 x_i$. Note that the actual observed value of *Y*, y_i, will not be exactly equal to $E[Y_i]$ because of variability; for example, not all voter turnouts for a particular level of advertising expenditure will be the same. $E[Y_i]$ is just the average value. It has a linear relationship with the value of *x*, as specified by the model.

The other part of the model is $\text{var}[Y_i]$ and $\text{cov}[Y_i, Y_j]$. This specifies that the variability of observations about $E[Y_i]$ is the same for all observations from the population, even though the $E[Y_i]$ changes as the value of *x* changes. Suppose I compute the variance of all observations in the population with a particular value of *x*. If I compare this variance with one similarly computed for another value of *x*, the model implies that the two variances should be the same. It is also specified that the covariance between two observations is zero. This implies that value of one of *Y* is not related to any other value of *Y*.

The model is illustrated in the following diagram. The thick line represents the relationship between $E[Y_i]$ and *x*. The parallel thin lines represent the standard deviation ($= \sqrt{\text{variance}}$) or the average deviation from $E[y_i]$.



The scatter diagram for Turnout versus Expend is as follows:



Does it look like the model will describe this situation? ■

I.C Model selection

Generally, we want to determine the model that best describes the data. To do this we generally obtain estimates of our parameters under several alternative models and use these in deciding which model to use to describe the data. The choice of models is often made using an analysis of variance (ANOVA).

a) Obtaining parameter estimates

Estimators of the parameters $\boldsymbol{\theta}$ in the expectation model are obtained using the least squares or maximum likelihood criteria — they are equivalent in the context of linear models. Here we will establish the least squares estimators. Also, an estimator of σ^2 is obtained from the ANOVA described in the next section.

Definition I.5: Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}$ is an $n \times q$ matrix with $n \geq q$, $\boldsymbol{\theta}$ is a $q \times 1$ vector of unknown parameters, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}_n$. The **ordinary least squares (OLS) estimator** of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that minimizes $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i^2$. ■

Note that $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ is of the form described in property 9 of lemma I.1 and is a scalar that is the sum of squares of the elements of $\boldsymbol{\varepsilon}$ or the sum of squares of the "errors".

The following theorem derives the ordinary least squares estimators for the full rank case. But what does full rank mean? The following definition tells us.

Definition I.6: The **rank** of an $n \times q$ matrix $\mathbf{A}$ with $n \geq q$ is the number of linearly independent columns of the matrix. The matrix is said to be of **full rank**, or rank q, if, none of the columns in the matrix can be written as a linear combination of the other columns. ■

Example I.1 House price (continued)

For this example the **X** matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 10 & 2 \\ 1 & 20 & 3 \end{bmatrix}$$

It is rank 3 and is full rank as none of the 3 columns can be written as a linear combination of the other two.

On the other hand the following matrices are of rank 2 as the second columns are 5×(3) and 5×(3) – 9×(1), respectively:

$$\mathbf{X} = \begin{bmatrix} 1 & 5 & 1 \\ 1 & 5 & 1 \\ 1 & 10 & 2 \\ 1 & 10 & 2 \\ 1 & 15 & 3 \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} 1 & -4 & 1 \\ 1 & -4 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 2 \\ 1 & 6 & 3 \end{bmatrix}$$

■

Theorem I.4: Let $\mathbf{Y} = \mathbf{X\theta} + \boldsymbol{\varepsilon}$ where **X** is an $n \times q$ matrix of full rank and $n \geq q$, $\boldsymbol{\theta}$ is a $q \times 1$ vector of unknown parameters, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors with mean **0** and variance $\sigma^2 \mathbf{I}_n$. The ordinary least squares estimator of $\boldsymbol{\theta}$ is denoted by $\hat{\boldsymbol{\theta}}$ and is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Proof: The vector of errors $\boldsymbol{\varepsilon}$ can be written as $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X\theta}$ and hence

$$\begin{aligned} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= (\mathbf{Y} - \mathbf{X\theta})'(\mathbf{Y} - \mathbf{X\theta}) \\ &= (\mathbf{Y}' - \boldsymbol{\theta}'\mathbf{X}')(\mathbf{Y} - \mathbf{X\theta}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X\theta} - \boldsymbol{\theta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\theta}'\mathbf{X}'\mathbf{X\theta} \end{aligned}$$

Since $\boldsymbol{\theta}'\mathbf{X}'\mathbf{Y}$ is 1×1, as noted in lemma I.1 it is symmetric yielding $\boldsymbol{\theta}'\mathbf{X}'\mathbf{Y} = (\boldsymbol{\theta}'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X\theta}$. Hence,

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X\theta} + \boldsymbol{\theta}'\mathbf{X}'\mathbf{X\theta}$$

To minimize $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ as a function of $\boldsymbol{\theta}$, this expression is differentiated with respect to $\boldsymbol{\theta}$, the derivative set equal to zero and the resulting equations solved for $\boldsymbol{\theta}$.

Note that for **a** an $n \times 1$ vector of constants, **A** an $n \times n$ matrix of constants and **z** an $n \times 1$ vector of variables, then $\mathbf{a}'\mathbf{z}$, $\mathbf{z}'\mathbf{z}$ and $\mathbf{z}'\mathbf{A}\mathbf{z}$ are all scalars that are functions of **z**. For any scalar, *u* say, that is a function of **z** we can take *n* partial derivatives of *u*, one with respect to each of the variables z_i. Let the column vector containing these be

$$\frac{\partial u}{\partial \mathbf{z}} = \begin{bmatrix} \partial u/\partial z_1 \\ \partial u/\partial z_2 \\ \vdots \\ \partial u/\partial z_n \end{bmatrix}$$

It can be shown that:

1. for $u = \mathbf{a}'\mathbf{z}$, $\partial u/\partial \mathbf{z} = \mathbf{a}$;
2. for $u = \mathbf{z}'\mathbf{z}$, $\partial u/\partial \mathbf{z} = 2\mathbf{z}$;
3. for $u = \mathbf{z}'\mathbf{A}\mathbf{z}$, $\partial u/\partial \mathbf{z} = \mathbf{A}\mathbf{z} + \mathbf{A}'\mathbf{z}$.

Now we require,

$$\frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \boldsymbol{\theta}} = \frac{\partial \left(\mathbf{Y}'\mathbf{Y} - 2(\mathbf{Y}'\mathbf{X})\boldsymbol{\theta} + \boldsymbol{\theta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}$$

and it is clear that the first term in the numerator does not depend on $\boldsymbol{\theta}$, that the second term is a scalar function of $\boldsymbol{\theta}$ of the first form above, and that the third term is a scalar function of $\boldsymbol{\theta}$ of the third form above.

Consequently,

$$\begin{aligned} \frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \boldsymbol{\theta}} &= -2(\mathbf{X}'\mathbf{Y}) + (\mathbf{X}'\mathbf{X})\boldsymbol{\theta} + (\mathbf{X}'\mathbf{X})'\boldsymbol{\theta} \\ &= -2(\mathbf{X}'\mathbf{Y}) + 2(\mathbf{X}'\mathbf{X})\boldsymbol{\theta} \end{aligned}$$

Setting this derivative to zero and substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ we obtain

$$-2(\mathbf{X}'\mathbf{Y}) + 2(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}} = 0 \text{ or } (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{Y}$$

These latter equations are called the **normal equations**. Now it can be shown that $rank(\mathbf{A}) = rank(\mathbf{A}') = rank(\mathbf{A}'\mathbf{A})$ so that $rank(\mathbf{X}'\mathbf{X}) = q$ since we have assumed **X** is of full rank. Hence, $(\mathbf{X}'\mathbf{X})^{-1}$ exists and we multiply both sides of the normal equations by it to obtain the ordinary least squares estimator of $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

as claimed. ■

For a particular example, we will have an observed vector **y** and this is substituted into the estimator to yield the estimate for that example.

Once you have estimates $\hat{\boldsymbol{\theta}}$ you can form the fitted values and residuals for the model.

Definition I.7: The estimator of the **expected values** for the expectation model $E[\mathbf{Y}] = \boldsymbol{\psi} = \mathbf{X}\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\psi}} = \mathbf{X}\hat{\boldsymbol{\theta}}$$

The estimates of Ψ for a particular observed **y** are called the **fitted values**. They are computed by substituting the values of the estimates of θ and the explanatory variables into the fitted equation. ■

Definition I.8: The estimator of the errors for the expectation model $E[\mathbf{Y}] = \boldsymbol{\psi} = \mathbf{X}\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{Y} - \hat{\boldsymbol{\psi}}$$

and so the estimates, the **residuals**, are computed by subtracting the fitted values from the observed values of the response variable. ■

Now it can be shown that $\hat{\boldsymbol{\psi}} = \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{Q}_{\mathrm{M}}\mathbf{Y}$, where $\mathbf{Q}_{\mathrm{M}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is an $n \times n$ projection matrix with the property that it is symmetric and idempotent.

Definition I.9: A matrix **E** is idempotent if $\mathbf{E}^2 = \mathbf{E}$. ■

Given that **X** is an $n \times q$ matrix,

- then $\mathbf{Q}_{\mathrm{M}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$
- is the product of $n \times q$, $q \times q$ and $q \times n$ matrices
- with the result that it is an $n \times n$ matrix.

Clearly the product of the $n \times n$ matrix $\mathbf{Q}_{\mathrm{M}}$ and the $n \times 1$ vector **Y** is an $n \times 1$ vector. So the estimator of the fitted values is a linear combination of the elements of the random vector for the response variable, **Y**.

Given the expression for the estimator of the expected values, the estimator of the errors is given by

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}} \\ &= \mathbf{Y} - \mathbf{Q}_{\mathrm{M}}\mathbf{Y} \\ &= \left(\mathbf{I}_n - \mathbf{Q}_{\mathrm{M}}\right)\mathbf{Y} \\ &= \mathbf{Q}_{\mathrm{R}}\mathbf{Y} \end{aligned}.$$

Again the result is an $n \times 1$ vector.

Hence the fitted values and residuals are given by $\hat{\boldsymbol{\psi}} = \mathbf{Q}_{\mathrm{M}}\mathbf{y}$ and $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\boldsymbol{\psi}} = \mathbf{Q}_{\mathrm{R}}\mathbf{y}$, respectively

Theorem I.5: Given that the matrix **E** is symmetric and idempotent, then the matrix $\mathbf{R} = \mathbf{I} - \mathbf{E}$ is also symmetric and idempotent. In addition, $\mathbf{ER} = \mathbf{RE} = \mathbf{0}$.

Proof: left as an exercise for you. ■

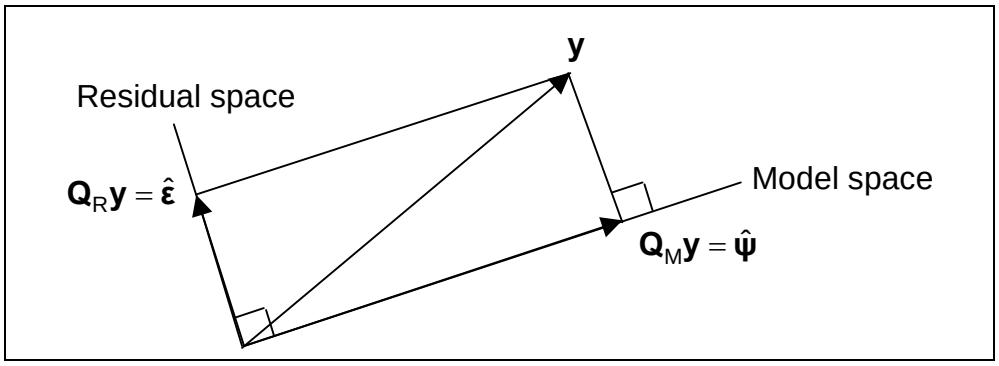
Application of this theorem to the regression situation leads us to conclude that $\mathbf{Q}_{\mathrm{R}}$ is symmetric and idempotent with $\mathbf{Q}_{\mathrm{R}}\mathbf{Q}_{\mathrm{M}} = \mathbf{Q}_{\mathrm{M}}\mathbf{Q}_{\mathrm{R}} = \mathbf{0}$.

All of this can be viewed as the orthogonal projection of vectors onto subspaces. The **observation vector y** is viewed as a vector in *n*-space and this space is called the **data space**.

Then the **X** matrix, with *q* linearly independent columns, determines a *q*-dimensional subspace of the data space — this space is called the **model (sub)space**. The fitted values, $\mathbf{X}\hat{\boldsymbol{\theta}}$, are the orthogonal projection of the observation vector into the model space. The orthogonal projection is achieved using the idempotent, or projection matrix, $\mathbf{Q}_M$.

The residuals are then the projection of the observation vector into the **residual subspace**, the subspace of the data space orthogonal to the model space — the matrix that projects onto the residual subspace is $\mathbf{Q}_R$. That $\mathbf{Q}_R\mathbf{Q}_M = \mathbf{Q}_M\mathbf{Q}_R = \mathbf{0}$ reflects that the two subspaces are orthogonal.

The situation for the estimates is represented in the following diagram.



The projection is orthogonal because there is a right angle between the fitted-values vector and the line from the data vector to the fitted-values vector.

Geometrically speaking, it is obvious why $\mathbf{Q}_M^2 = \mathbf{Q}_M$.

- Once you have projected **y** into the model subspace and obtained $\mathbf{Q}_M\mathbf{y}$, it is in the model subspace.
- Applying $\mathbf{Q}_M$ to the fitted values, that is to $\mathbf{Q}_M\mathbf{y}$, will have no effect because they are already in the model subspace.
- Clearly, $\mathbf{Q}_M^2\mathbf{y} = \mathbf{Q}_M\left(\mathbf{Q}_M\mathbf{y}\right) = \mathbf{Q}_M\mathbf{y}$.

A similar argument applies to $\mathbf{Q}_R$.

Also, it should be clear why $\mathbf{Q}_R\mathbf{Q}_M = \mathbf{0}$ — look at the result of applying $\mathbf{Q}_M$ to **y** and then applying $\mathbf{Q}_R$ to the result.

Example I.3 Single sample

Suppose that a single sample of 3 observations has been obtained. The linear model we propose for this data is that

$$E\left[\mathbf{Y}\right] = \mathbf{X}_G\mu = \begin{bmatrix}1\\1\\1\end{bmatrix}\mu = \mathbf{1}_3\mu \text{ and } \operatorname{var}\left[\mathbf{Y}\right] = \sigma^2\mathbf{I}_n .$$

or, for an individual observation,

$$Y_i = \mu + \varepsilon_i \text{ with } \text{var}\left[Y_i\right] = \sigma^2 \text{ and } \text{cov}\left[Y_i, Y_j\right] = 0,\ i \neq j\,.$$

That is, the value for an observation is made up of the population mean plus a particular deviation from the population mean for that observation.

In this case $\mathbf{Q}_{\text{M}}$, a 3×3 matrix, is rather simple as

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \mathbf{1}_3$$

so that

$$\begin{aligned} \mathbf{Q}_{\text{M}} &= \mathbf{X}_{\text{G}}\left(\mathbf{X}'_{\text{G}}\mathbf{X}_{\text{G}}\right)^{-1}\mathbf{X}'_{\text{G}} \\ &= \mathbf{1}_3\left(\mathbf{1}'_3\mathbf{1}_3\right)^{-1}\mathbf{1}'_3 \\ &= \mathbf{1}_3\left(3\right)^{-1}\mathbf{1}'_3 \\ &= \tfrac{1}{3}\mathbf{1}_3\mathbf{1}'_3 \\ &= \tfrac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \\ &= \tfrac{1}{3}\mathbf{J}_3 \end{aligned}$$

and

$$\hat{\boldsymbol{\psi}} = \mathbf{Q}_{\text{M}}\mathbf{Y} = \tfrac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\mathbf{Y} = \begin{bmatrix} \bar{Y} \\ \bar{Y} \\ \bar{Y} \end{bmatrix}.$$

That is, in this case, $\mathbf{Q}_{\text{M}}$ is the matrix that replaces each observation with the grand mean of all the observations. We will use $\mathbf{Q}_{\text{G}}$ for it and $\bar{\mathbf{G}}$ to denote the vector of grand means throughout this course. Hence, $\bar{\mathbf{G}} = \mathbf{1}_3\bar{Y}$ and $\bar{\mathbf{g}} = \mathbf{1}_3\bar{y}$.

Note that the estimator of μ in our model is the mean of the elements of $\mathbf{Y}$, $\bar{Y}$, and the estimate is the mean of the observations, $\bar{y}$.

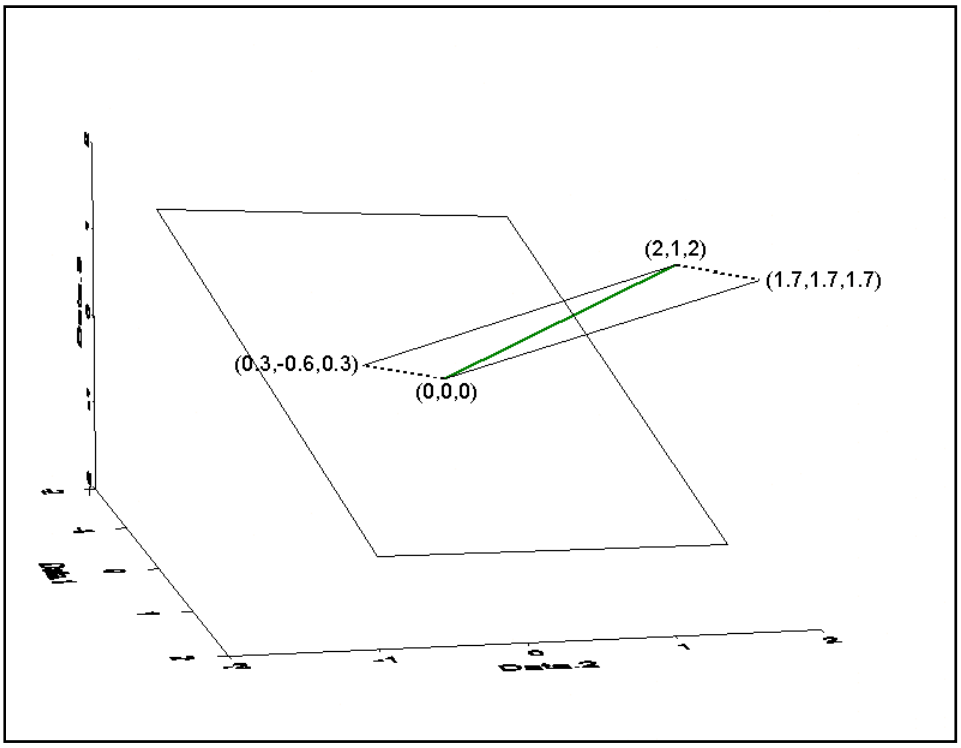
Suppose that $\mathbf{y}' = \left(2,1,2\right)$.

Then $\bar{y} = 5/3 = 1.67$ and fitting the model $E\left[\mathbf{Y}\right] = \mathbf{1}_n\mu$ results in

- fitted values $\hat{\boldsymbol{\psi}}' = \left(1.67,1.67,1.67\right)$ and
- residuals $\hat{\boldsymbol{\varepsilon}}' = \left(0.33,-0.66,0.33\right)$.

Now as there are 3 observations, our data space is 3-space and we can illustrate this in a 3D-plot.

A plot of these vectors is shown in the following diagram (the first data point is plotted on the axis coming out of the figure, the second on the axis going across and the third on the axis going up).



The model subspace is the equiangular line and the residual subspace is the plane orthogonal to the model subspace. The vertical side of the rectangle is the fitted vector and the dotted line is the residual vector. ■

b) Regression analysis of variance

An analysis of variance is used to compare potential models. In the case of the regression model, it is common to want to choose between two expectation models, one which is a subset of the other.

Testing all expectation parameters are zero

The simplest, although not necessarily the most useful, situation is where one compares the expectation models

- $E[Y_i] = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$ and
- $E[Y_i] = 0$.

So we first state the null and alternative hypothesis for the hypothesis test.

H_0: $\boldsymbol{\theta} = 0$ (equivalent to $E[Y_i] = 0$)

H_1: $\boldsymbol{\theta} \neq 0$ (equivalent to $E[Y_i] = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$)

The test statistic is computed using an analysis of variance table.

Source	DF	SSq	MSq	F	p
Model_0	0	$\mathbf{0}'\mathbf{0}$			
$\text{Model}_1 - \text{Model}_0$ (Model)	q	$\hat{\boldsymbol{\psi}}'\hat{\boldsymbol{\psi}} - \mathbf{0}'\mathbf{0}$	$\frac{\hat{\boldsymbol{\psi}}'\hat{\boldsymbol{\psi}}}{q}\left(=s_{\text{M}}^2\right)$	$\frac{s_{\text{M}}^2}{s_{\text{R}}^2}$	$\Pr\left\{F_{q,n-q} \geq F_O\right\}$
Residual	$n-q$	$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$	$\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-q}\left(=s_{\text{R}}^2\right)$		
Total	n	$\mathbf{Y}'\mathbf{Y}$			

Generally an analysis of variance involving the comparison of two models involves the sum of squares of the estimators of the expected values for the null model and the difference between the sums of squares of the estimators of the expected values for the two models. In this case, the estimator for null model are all 0 and so the difference in the sums of squares is equal to the sum of squares of the estimator of the expected values of the alternative model. We could leave Model_0 out of the table altogether.

This table involves two parallel identities: the degrees-of-freedom and sums-of-squares identities. It is obvious that Total df = Model df + Residual df. What is not so clear is that Total SSq = Model SSq + Residual SSq; nonetheless it is true as is illustrated by the geometrical interpretation.

Note the use of s^2, the symbol for the variance, for a mean square. This is because mean squares are indeed variances, being the ratio of a sum of squares to its degrees of freedom. It is an accident of history that the term mean square, rather than variance, is used in the context of ANOVA.

If the p-value is less than the significance level, α, the null hypothesis is rejected. Usually, $\alpha = 0.05$.

Note that a vector transposed multiplied by the vector itself is a sum of squares (see property 9 of lemma I.1). For example,

$$\mathbf{Y}'\mathbf{Y} = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sum_{i=1}^{n} Y_i^2$$

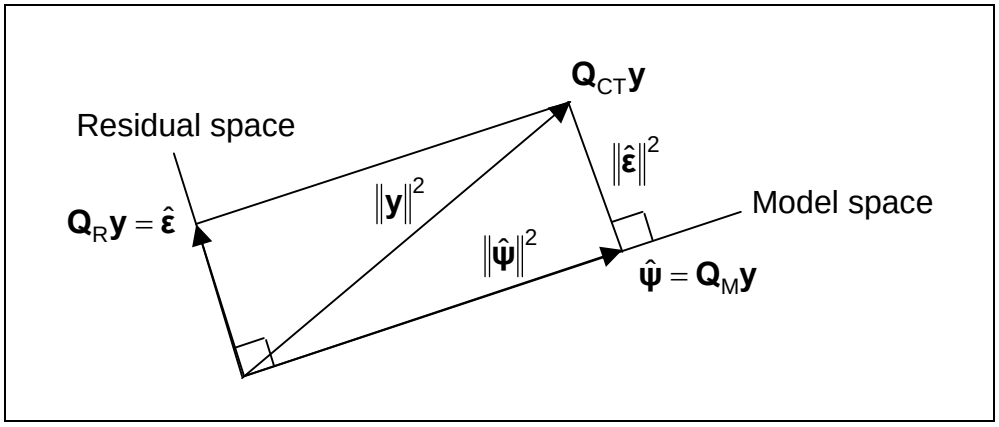
So the estimator of an SSq or sum of squares are exactly that — the sum of squares of:

Model (or Regression) SSq: the estimator of the expected values under the alternative model.

Residual SSq: the estimator of the errors obtained by subtracting the expected values under the alternative model from the random vector **Y**.

Total SSq: the elements of the random vector **Y**.

Now the squared length of a vector is equal to the a sum of squares of its elements so that, diagrammatically, we have for the estimates:



Now from Pythagoras' theorem we have that $\|\mathbf{y}\|^2 = \|\hat{\boldsymbol{\psi}}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2$. However, this is seen to be equivalent to $\mathbf{y}'\mathbf{y} = \hat{\boldsymbol{\psi}}'\hat{\boldsymbol{\psi}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ which proves the sum of squares identity noted earlier.

Example I.3 Single sample (continued)

Recalling that $\mathbf{y}' = (2,1,2)$, $\hat{\boldsymbol{\psi}}' = (1.67,1.67,1.67)$ and $\hat{\boldsymbol{\varepsilon}}' = (0.33,-0.66,0.33)$, it is easy to verify that the squared lengths, or sums of squares, are 9, 8.33 and 0.67 for total, fitted and residual, respectively.

We note that because the data is very close to the fitted line, there is only a small vector in the residual space and its squared length is small relative to that for the fitted vector. However, we should take into account that the fitted values involves only one value and so has only 1 degree of freedom whereas the residuals has two independent values and so 2 degrees of freedom. Dividing each sum of squares by it degrees of freedom, to yield mean squares, results in 8.33 and 0.33. The difference is only more marked. ■

Example I.1 House price (continued)

For this example, using the computer we find that

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} 33.06 \\ -0.189 \\ 10.178 \end{bmatrix}$$

The estimated expected value is given by

$$\widehat{E[Y]} = 33.0626 - 0.1897x_1 + 10.7182x_2 .$$

In this case $\mathbf{Q}_M$, a 5×5 projection matrix, is somewhat more complicated. Using R it can be shown that

$$\mathbf{Q}_M = \begin{bmatrix} 0.448 & 0.359 & 0.249 & 0.138 & 0.193 \\ 0.359 & 0.683 & -0.245 & 0.160 & 0.042 \\ 0.249 & -0.245 & 0.805 & 0.188 & 0.004 \\ 0.138 & 0.160 & 0.188 & 0.215 & 0.298 \\ 0.193 & 0.042 & 0.004 & 0.298 & 0.849 \end{bmatrix}$$

We can obtain the fitted values either by substituting the values of the two explanatory variables into the fitted equation or by applying $\mathbf{Q}_M$ to $\mathbf{y}$. The fitted values and the residuals are as follows:

	Observations $(=\mathbf{y})$	Fitted values $(\hat{\boldsymbol{\psi}} = \mathbf{Q}_M\mathbf{y})$	Residuals $\left(\begin{matrix}\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{Q}_M\mathbf{y} \\ = \mathbf{Q}_R\mathbf{y}\end{matrix}\right)$
	50	43.59116	6.408840
	40	42.83241	-2.832413
	52	53.55064	-1.550645
	47	52.60221	-5.602210
	65	61.42357	3.576427
SSq		13142.32	95.67588

Note that the Observations are equal to sums of Fitted values and the Residuals and that the sum of the last two sums of squares is approximately equal to the Total sum of squares.

The analysis of variance table for the example is:

Source	**DF**	**SSq**	**MSq**	**F**	**p**
Regression	3	13142.32	4380.78	91.58	0.0108
Residual	2	95.68	47.84		
Total	5	13238.00			

Note that the p-value is obtained using R.

As the p-value is less than 0.05, the null hypothesis is rejected. The expectation model $E[Y_i] = 0$ does not provide as good a description of the data as the model $E[Y_i] = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$. ■

Testing that a subset of the expectation parameters are zero

A more useful test involves testing that just some of the θs are zero. For example, in multiple linear regression you might want to choose between the expectation models $E[Y_i] = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$ and $E[Y_i] = \theta_0$.

Again, we first state the null and alternative hypothesis for the hypothesis test.

H_0: $\theta_1 = \theta_2 = 0$ (equivalent to $E[Y_i] = \theta_0$)

H_1: $\theta_1, \theta_2 \neq 0$ (equivalent to $E[Y_i] = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$)

The test statistic is computed using an analysis of variance table.

Source	**DF**	**SSq**	**MSq**	**F**	**p**
Model_0	1	$\bar{\mathbf{G}}'\bar{\mathbf{G}}$			
$\text{Model}_1 - \text{Model}_0$	$q-1$	$\hat{\boldsymbol{\psi}}_1'\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}'\bar{\mathbf{G}}$	s_M^2	$\frac{s_M^2}{s_R^2}$	$\Pr\{F_{1,n-2} \geq F_O\}$
Residual	$n-q$	$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$	s_R^2		
Total	n	$\mathbf{Y}'\mathbf{Y}$			

$s_M^2 = \left(\hat{\boldsymbol{\psi}}_1'\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}'\bar{\mathbf{G}}\right)/(q-1)$ and $s_R^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-q)$ where $\bar{\mathbf{G}} = \mathbf{1}_n\bar{Y}$

Now the null model is not so trivial as when testing all parameters are zero. In this case the null model is equal to grand mean model considered in Example I.3, *Single sample*. We showed there that the estimator of the expected values vector was $\bar{\mathbf{G}}$, all of whose elements are equal to the grand mean estimator.

Note that the difference in the sum of squares does not look like a sum of squares, but a difference in sums of squares. However, can show that $\hat{\boldsymbol{\psi}}_1'\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}'\bar{\mathbf{G}} = \left(\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}\right)'\left(\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}\right)$ — that is it makes no difference whether the difference in the estimators of the expected values is obtained and the sum of squares of the difference computed, or the difference in the sum of squares of the two sets of estimators of the expected values is obtained. This only works for two models, one of which is a subset of the other.

Also, it is so unusual to test a hypothesis about a model that does not include the intercept term that the sum of squares for the model involving only it is usually subtracted out of the analysis of variance. Thus the total of the sums of squares becomes the **Corrected Total sums of squares**. Again, one can either subtract the grand mean from the observations and form the sum of squares or subtract the sum of squares for the grand mean model from the uncorrected total sum of squares — that is $\mathbf{Y}'\mathbf{Y} - \bar{\mathbf{G}}'\bar{\mathbf{G}} = \left(\mathbf{Y} - \bar{\mathbf{G}}\right)'\left(\mathbf{Y} - \bar{\mathbf{G}}\right)$. We prefer the later as it is a sum of squares, rather than the former that is a difference of sums of squares.

Source	DF	SSq	MSq	F	p
$\text{Model}_1 - \text{Model}_0$ (Model)	$q-1$	$\left(\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}\right)'\left(\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}\right)$	s_{M}^2	$\frac{s_{\text{M}}^2}{s_{\text{R}}^2}$	$\Pr\left\{F_{1n-2} \geq F_O\right\}$
Residual	$n-q$	$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$	s_{R}^2		
(Corrected) Total	$n-1$	$\left(\mathbf{Y} - \bar{\mathbf{G}}\right)'\left(\mathbf{Y} - \bar{\mathbf{G}}\right)$			

$s_{\text{M}}^2 = \left(\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}\right)'\left(\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{G}}\right)\Big/(q-1)$ and $s_{\text{R}}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n-q)$

The SSqs in this analysis are the sum of squares of the following quantities:

Model SSq: the differences between the estimators of the expected values for the two models in the hypotheses.

Residual SSq: the estimators of the errors obtained by subtracting the estimators of the expected values under the alternative model from the random vector **Y**.

(Corrected) Total SSq: the deviations from the grand mean obtained by subtracting the grand mean estimator from the random vector **Y**.

Example I.1 House price (continued)

Taking the previously computed fitted values and residuals and subtracting $\bar{\mathbf{g}} = 50.8\,\mathbf{1}_5$ from the response variable and the fitted values we obtain:

	Observations $(\mathbf{y})$	Deviations $(\mathbf{y} - \bar{\mathbf{g}})$	Fitted values $(\hat{\boldsymbol{\psi}}_1)$	Model differences $(\hat{\boldsymbol{\psi}}_1 - \bar{\mathbf{g}})$	Residuals $(\mathbf{y} - \hat{\boldsymbol{\psi}}_1)$
	50		43.59116	-7.20884	6.408840
	40		42.83241	-7.96759	-2.832413
	52		53.55064	2.75064	-1.550645
	47		52.60221	1.80221	-5.602210
	65		61.42357	10.62357	3.576427
SSq	13238		13142.32	239.12409	95.67588

Note that the Deviations are equal to sum of the Model differences and the Residuals and that the sum of the last two sums of squares is approximately equal to the Deviations sum of squares.

The analysis of variance table for the example is:

Source	DF	SSq	MSq	F	p
Regression	2	239.124	119.562	2.50	0.2857
Residual	2	95.676	47.838		
Total	4	334.800			

As the p-value is greater than 0.05, the null hypothesis cannot be rejected. The expectation model $E[Y_i] = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$ does not describe the data any better than the model $E[Y_i] = \theta_0$. As the latter model is simpler, it will be adopted as the model that best describes the data ■

I.D Summary

In this chapter, we have

- looked at the definition and meaning of the expectation and variance of a random variable;
- investigated the expectation of a linear function of a random variable;
- derived the probability distribution function to be used to describe a random sample;
- examined the expectation of a linear combination of functions of random variables;
- formulated linear regression models and described how estimates of the parameters in these models are obtained using matrices;
- provided hypothesis tests for testing hypotheses about the parameters using an analyses of variance involving sums of squares of various quantities.

I.E Exercises

I.1 Suppose that Y is a random variable that represents the actual contents of a 1-lb can of coffee. The model proposed for the distribution of Y is the uniform distribution over the interval $[15.5,17.0]$

$$f(y) = \frac{1}{1.5}, \quad 15.5 \le y \le 17.0$$

a) What is the probability that a can will contain less than 16oz?
b) Find the population mean and standard deviation for these cans.

I.2 Verify that $\mathbf{V} = E\left[\left(\mathbf{Y} - E[\mathbf{Y}]\right)\left(\mathbf{Y} - E[\mathbf{Y}]\right)'\right]$ is equivalent to the following expression for $\mathbf{V}$ by obtaining an expression for the *ij*th element of $E\left[\left(\mathbf{Y} - E[\mathbf{Y}]\right)\left(\mathbf{Y} - E[\mathbf{Y}]\right)'\right]$.

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1j} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2i} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{1j} & \sigma_{2i} & \cdots & \sigma_i^2 & \cdots & \sigma_{jn} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_{in} & \cdots & \sigma_n^2 \end{bmatrix}$$

I.3 Prove that $\frac{1}{3}\mathbf{J}_3$ is idempotent, where $\mathbf{J}_3$ is the 3×3 matrix all of whose elements are equal to 1.

I.4 Let x denote the number of years of formal education and let Y denote an individual's income at age 30. Assume that simple linear regression is applicable and consider this data:

Formal education (years)	Income ($000)
8	8
12	15
14	16
16	20
16	25
20	40

a) Write down $\mathbf{y}$, $\mathbf{X}$ and $\boldsymbol{\theta}$ for this data.

b) Use the R function `lm` to find $\hat{\boldsymbol{\theta}}$. What is the equation for the estimated expected value?

c) You are given that

d) $$\mathbf{Q}_M = \begin{bmatrix} 0.648 & 0.344 & 0.192 & 0.040 & 0.040 & -0.264 \\ 0.344 & 0.232 & 0.176 & 0.120 & 0.120 & 0.008 \\ 0.192 & 0.176 & 0.168 & 0.160 & 0.160 & 0.144 \\ 0.040 & 0.120 & 0.160 & 0.200 & 0.200 & 0.280 \\ 0.040 & 0.120 & 0.160 & 0.200 & 0.200 & 0.280 \\ -0.264 & 0.008 & 0.144 & 0.280 & 0.280 & 0.552 \end{bmatrix}$$

e) Compute the fitted values by calculating $\mathbf{Q}_M\mathbf{y}$.

f) Use the equation for the estimated expected value to verify the first fitted value.

g) Use the fitted values to compute the residuals.

h) Use the R function `aov` to obtain the ANOVA table for testing that the slope is zero given that the intercept is in the model.
What is the corrected total SSq for this analysis?
Verify that the Residual SSq is the sum of the squares of the residuals.

i) What model best describes this data?