THE DESIGN AND MIXED-MODEL ANALYSIS OF EXPERIMENTS

II. Statistical inference in regression

II.A The linear model in regression

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.

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

All but the last two are linear in the θ_is.

We would conduct a study in which n (≥ 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 in the above model.

It is also usual to make further assumptions about ε_is: $E[\varepsilon_i] = 0$, $\text{var}[\varepsilon_i] = \sigma^2$ 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 quantity not encountered before: the covariance between two different errors.

Definition II.1: The covariance of two random variables, X and Y, is defined to be

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

■

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:

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

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 II.2: 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 $\text{cov}\left[Y_i, Y_j\right]=\sigma_{ij}\ (=\sigma_{ji})$. Then, the **random vector** is

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

The **expectation vector** 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

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

■

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

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

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 $E[\mathbf{Y}]$ and $\operatorname{var}[\mathbf{Y}]$. These alternative expressions can be obtained by substituting the model into $E[Y_i]$, $\operatorname{var}[Y_i]$ and $\operatorname{cov}\left[Y_i, Y_j\right]$. Thus,

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

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

and

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

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

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

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

Example II.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 $E[\varepsilon_i] = 0$, $\mathrm{var}[\varepsilon_i] = \sigma^2$ and $\mathrm{cov}\left[\varepsilon_i, \varepsilon_j\right] = 0,\ i \neq j$,

or, equivalently,

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

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

In matrix terms, the model is:

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

or, equivalently,

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

where

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

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

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

Our task is to find estimators of $\boldsymbol{\theta}$.

Example II.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 variable number subscript so that our proposed model is:

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

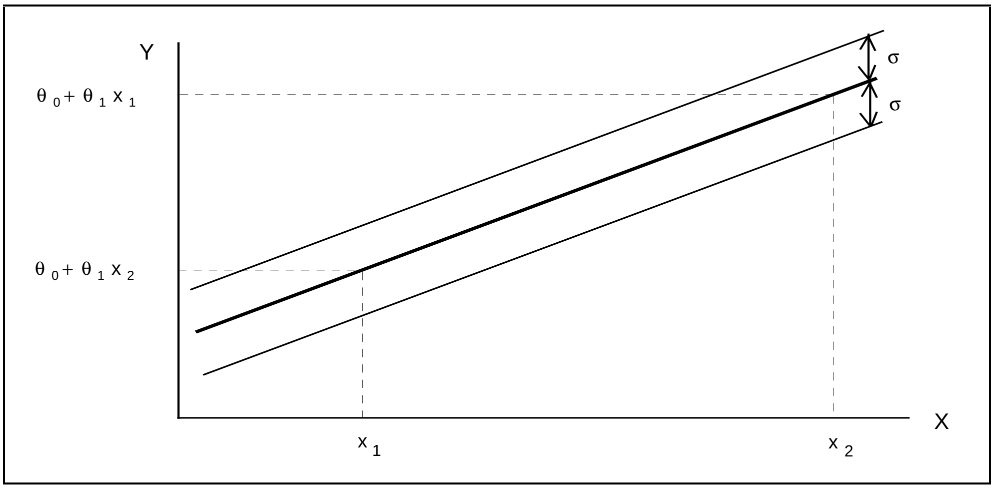
with $\operatorname{var}\left[Y_i\right] = \sigma^2$ and $\operatorname{cov}\left[Y_i, Y_j\right] = 0,\ i \neq j$,

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?

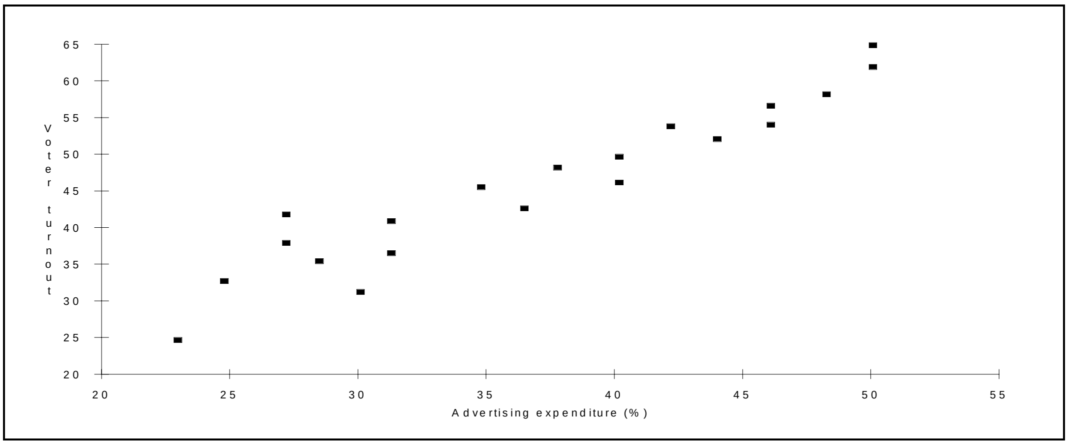
$E\left[Y_i\right]$ means the population average value or population mean response. The equation given above for $E\left[Y_i\right]$ implies that the average value of Y, for individuals for which the value of x is x_i, is $\theta_0 + \theta_1 x_i$. Thus, if the value of x for a particular observation 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 $E\left[Y_i\right]$ because of variability; for example, not all voter turnouts for a particular level of advertising expenditure will be the same. $E\left[Y_i\right]$ 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 $\operatorname{var}\left[Y_i\right]$ and $\operatorname{cov}\left[Y_i, Y_j\right]$. This specifies that the variability of observations about $E\left[Y_i\right]$ is the same for all observations from the population, even though the $E\left[Y_i\right]$ 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 variance 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?

II.B Least squares estimation of the expectation model parameters

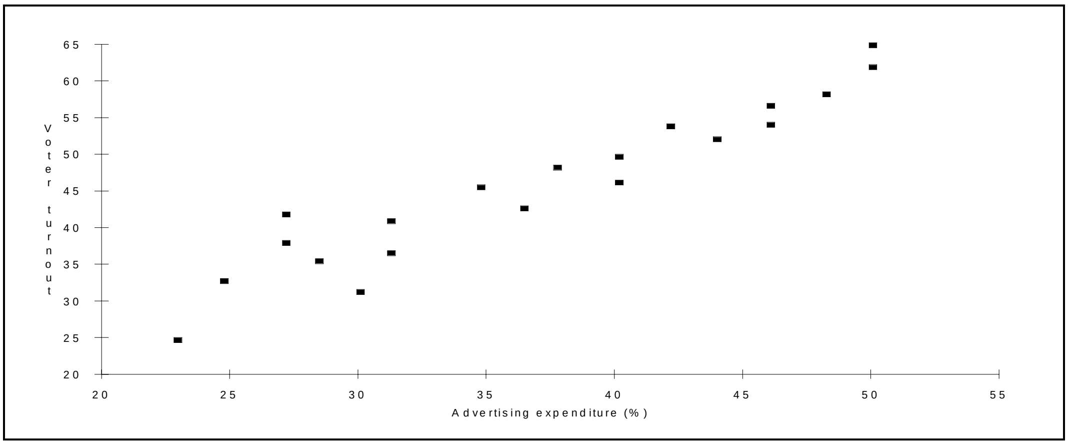
At this stage, our principal problem is to estimate the values of our population parameters θ_0 and θ_1. In order to do this we want to establish estimators for them. There are several different methods for doing this. A common method is the method of least squares.

Definition II.3: Let $\mathbf{Y}=\mathbf{X\theta}+\boldsymbol{\varepsilon}$ where $\mathbf{X}$ is an $n\times q$ matrix of full rank, $\boldsymbol{\theta}$ is a $q\times1$ vector of unknown parameters, $\boldsymbol{\varepsilon}$ is an $n\times1$ vector of errors with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}_n$, $q=p+1$ and $n\geq q$. The **least squares estimator** of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that minimizes $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}=\sum_{i=1}^{n}\varepsilon_i^2$, the sum of squares of the "errors". ■

Note that $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ is a scalar.

Example II.2 Voter Turnout (continued)

What we are wanting to do can be illustrated on the plot of Turnout versus Expend:



The aim is to put a straight line through the points such that the sum of squares of the vertical distances of the observations from the line is a minimum.

a) Least squares estimators

Theorem II.1: Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}$ is an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ 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 $\sigma^2\mathbf{I}_n$, $q = p + 1$ and $n \geq q$. The least squares estimator for $\boldsymbol{\theta}$ is denoted by $\hat{\boldsymbol{\theta}}$ and is given by

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

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

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

Since $\boldsymbol{\theta}'\mathbf{X}'\mathbf{Y}$ is 1×1, $\boldsymbol{\theta}'\mathbf{X}'\mathbf{Y} = \left(\boldsymbol{\theta}'\mathbf{X}'\mathbf{Y}\right)' = \mathbf{Y}'\mathbf{X}\boldsymbol{\theta}$ and so

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

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 we obtain

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

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 least squares estimators 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.

Example II.1 House price (continued)

For this example,

$$\mathbf{y}=\begin{bmatrix}50\\40\\52\\47\\65\end{bmatrix}\quad\text{and}\quad\mathbf{X}=\begin{bmatrix}1&1&1\\1&5&1\\1&5&2\\1&10&2\\1&20&3\end{bmatrix}$$

so that

$$\mathbf{X}'\mathbf{X}=\begin{bmatrix}5&41&9\\41&551&96\\9&96&19\end{bmatrix}\quad\text{and}\quad\mathbf{X}'\mathbf{y}=\begin{bmatrix}254\\2280\\483\end{bmatrix}$$

Using the computer we find that

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}=\begin{bmatrix}2.307551&0.1565378&-1.88398\\0.1565378&0.02578269&-0.20442\\-1.88398&-0.20442&1.977901\end{bmatrix}$$

so that the least squares estimates are given by

$$\begin{aligned}\hat{\boldsymbol{\theta}}&=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\\&=\begin{bmatrix}2.307551&0.1565378&-1.88398\\0.1565378&0.02578269&-0.20442\\-1.88398&-0.20442&1.977901\end{bmatrix}\begin{bmatrix}254\\2280\\483\end{bmatrix}\\&=\begin{bmatrix}33.06\\-0.189\\10.178\end{bmatrix}\end{aligned}$$

The estimated expected value is

$$\widehat{E\left[Y\right]}=33.06-0.189x_1+10.718x_2$$

Example II.2 Voter Turnout (continued)

For this example, $\mathbf{X}'\mathbf{X}=\begin{bmatrix}20&739.8\\739.8&28825.7\end{bmatrix}$ and $\mathbf{X}'\mathbf{y}=\begin{bmatrix}914.2\\35535.3\end{bmatrix}$.

To invert $\mathbf{X}'\mathbf{X}$ use the well-known result that the inverse of the 2 x 2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $\dfrac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

b) Properties of least squares estimators

Now, we have established the least squares estimators, but are they unbiased? That is, is $E\left[\hat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}$.

Theorem II.2: Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}$ is an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ 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 $\sigma^2 \mathbf{I}_n$, $q = p+1$ and $n \geq q$. The least squares estimator $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is an unbiased estimator for $\boldsymbol{\theta}$. Furthermore, $\text{var}\left[\hat{\boldsymbol{\theta}}\right] = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$

Proof: The proof uses the rules given in theorem II.3 for obtaining the expectation and variance of a random vector. It is left as an exercise for you. ■

Theorem II.3: Let $\mathbf{a}$ be a $k \times 1$ vector of constants, $\mathbf{A}$ a $m \times k$ matrix of constants and $\mathbf{Y}$ a $k \times 1$ vector of random variables or random vector.

1. $E[\mathbf{a}] = \mathbf{a}$ and $\text{var}[\mathbf{a}] = \mathbf{0}$;
2. $E[\mathbf{a}'\mathbf{Y}] = \mathbf{a}'E[\mathbf{Y}]$ and $\text{var}[\mathbf{a}'\mathbf{Y}] = \mathbf{a}'\text{var}[\mathbf{Y}]\mathbf{a}$;
3. $E[\mathbf{AY}] = \mathbf{A}E[\mathbf{Y}]$ and $\text{var}[\mathbf{AY}] = \mathbf{A}\,\text{var}[\mathbf{Y}]\mathbf{A}'$.

Proof: The proof of this theorem uses theorem I.6 that gives the expectation of a linear combination of functions of a multivariate random variable and definitions I.4, I.5 and II.2 that define the expectation and variance of random variables and vectors. ■

Gauss-Markoff theorem

Definition II.4: **Linear estimators** are estimators of the form $\mathbf{LY}$, where $\mathbf{L}$ is a matrix of real numbers. ■

The unbiased least squares estimator developed here is an example of a linear estimator with $\mathbf{L} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Unfortunately, unbiasedness does not guarantee uniqueness. There might be more than one set of unbiased estimators for $\boldsymbol{\theta}$. In the previous section we suggested that another desirable property of an estimator was that it be minimum variance. Theorem II.4, the Gauss-Markoff theorem, guarantees that among the class of linear unbiased estimators for $\hat{\boldsymbol{\theta}}$ the least squares estimator is the best in the sense that the variances of the estimator, $\text{var}\left(\hat{\boldsymbol{\theta}}\right)$, is minimized. For this reason the least squares estimator is called the BLUE (**b**est **l**inear **u**nbiased **e**stimator).

Theorem II.4: Let $\mathbf{Y} = \mathbf{X\theta} + \boldsymbol{\varepsilon}$ where $\mathbf{X}$ is an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ 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 $\sigma^2 \mathbf{I}_n$, $q = p+1$ and $n \geq q$. The least squares estimator $\hat{\boldsymbol{\theta}}$ is the best linear unbiased estimator for $\boldsymbol{\theta}$.

Proof: Let $\hat{\boldsymbol{\theta}}^*$ denote any other linear unbiased estimator for $\boldsymbol{\theta}$. Without loss of generality, this estimator can be written in the form

$$\hat{\boldsymbol{\theta}}^* = \left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' + \mathbf{B} \right] \mathbf{Y}$$

where $\mathbf{B}$ is $q \times n$ matrix of real numbers. Taking expectations,

$$\begin{aligned} E\left[\hat{\boldsymbol{\theta}}^*\right] &= \left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' + \mathbf{B} \right\} E[\mathbf{Y}] \\ &= \left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' + \mathbf{B} \right\} \mathbf{X\theta} \\ &= \left\{ \mathbf{I}_q + \mathbf{BX} \right\} \boldsymbol{\theta} \end{aligned}$$

Since $\hat{\boldsymbol{\theta}}^*$ is an unbiased estimator for $\boldsymbol{\theta}$, $E\left[\hat{\boldsymbol{\theta}}^*\right] = \boldsymbol{\theta}$ and so $\left\{ \mathbf{I}_q + \mathbf{BX} \right\} \boldsymbol{\theta} = \boldsymbol{\theta}$. Thus, $\mathbf{I}_q + \mathbf{BX} = \mathbf{I}_q$ and $\mathbf{BX} = \mathbf{0}$. By the rules for the variance given in theorem II.3,

$$\begin{aligned} \operatorname{var}\left[\hat{\boldsymbol{\theta}}^*\right] &= \operatorname{var}\left[\left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' + \mathbf{B} \right\} \mathbf{Y} \right] \\ &= \left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' + \mathbf{B} \right\} \sigma^2 \mathbf{I}_q \left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' + \mathbf{B} \right\}' \\ &= \sigma^2 \left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' + \mathbf{B} \right\} \left\{ \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{B}' \right\} \\ &= \sigma^2 \left\{ (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{B}' + \mathbf{BX} (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{BB}' \right\} \end{aligned}$$

Since $\mathbf{BX} = \mathbf{0}$ and $\mathbf{X}'\mathbf{B}' = \mathbf{0}$,

$$\begin{aligned} \operatorname{var}\left[\hat{\boldsymbol{\theta}}^*\right] &= \sigma^2 \left\{ (\mathbf{X}'\mathbf{X})^{-1} + \mathbf{BB}' \right\} \\ &= \operatorname{var}\left[\hat{\boldsymbol{\theta}}\right] + \sigma^2 \mathbf{BB}' \end{aligned}$$

Now $\operatorname{var}\left[\hat{\boldsymbol{\theta}}^*\right]$ is the variance matrix of the estimator and we only require to minimize the variance. That is, we want the diagonal elements of this matrix to be minimum. Now the *i*th entry of $\mathbf{BB}'$ is the sum of squares of the *i*th row of $\mathbf{B}$:

$$\sum_{j=1}^{n} b_{ij}^2 \geq 0,\ \ i=1,2,\ldots,n$$

Consequently, the entires of the main diagonal of $\text{var}\left[\hat{\boldsymbol{\theta}}^*\right]$ are minimized when $\mathbf{B}=\mathbf{0}$. In this case $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}$ and the least squares estimator $\hat{\boldsymbol{\theta}}$ is the minimum variance estimator for $\boldsymbol{\theta}$ as claimed. ■

Definition II.5: A **uniformly minimum variance unbiased estimator** (UMVUE) is one for that has the smallest variance amongst *all* unbiased estimators. ■

The difference between a BLUE and an UMVUE is that one can only guarantee that a BLUE is best amongst *linear* unbiased estimators. It can be proved that, provided *Y* is normally distributed, $\hat{\boldsymbol{\theta}}$ is a UMVUE.

c) Estimating linear functions of the parameters

Suppose we need to estimate some linear function of the parameters. Such a function can be written as ℓ'q where ℓ' is a $1\times q$ vector of real numbers. The logical estimator is $\ell'\hat{\boldsymbol{\theta}}$. Theorem II.5 extends theorem II.4 to this more general estimation problem.

Theorem II.5: Let $\mathbf{Y}=\mathbf{X}\boldsymbol{\theta}+\boldsymbol{\varepsilon}$ where $\mathbf{X}$ is an $n\times q$ matrix of full rank, $\boldsymbol{\theta}$ 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 $\sigma^2\mathbf{I}_n$, $q=p+1$ and $n\geq q$. Let ℓ' is a $1\times q$ vector of real numbers. The best linear unbiased estimator for $\ell'\boldsymbol{\theta}$ is $\ell'\hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ is the least squares estimator.

Proof: The proof of this theorem parallels that of theorem II.4. ■

One important use of this theorem is to find the estimated mean response.

Example II.1 House price (continued)

Suppose you want to estimate the average house price for a 15 year-old house with 2500 feet^2 of living space. We have the following estimated equation for the expected or average value:

$$\hat{E}[Y]=33.06-0.189x_1+10.718x_2$$

This equation provided us with the estimated mean response if we substitute values of x$_1$ and x$_2$ into it. In our case we are interested in $x_1=15$ and $x_2=2.5$. However, we can write the estimated mean response as a linear combination of the estimated parameters as follows:

$$\begin{aligned}\ell'\hat{\boldsymbol{\theta}} &= \begin{bmatrix}1 & 15 & 2.5\end{bmatrix}\begin{bmatrix}33.06\\ -0.189\\ 10.178\end{bmatrix}\\ &= 33.06 - 0.189(15) + 10.178(2.5)\\ &= 57.02\end{aligned}$$

The estimated selling price of such houses is \$57 020. Our theorem tells is that this estimate is BLU.

II.C Estimating the variance

As we stated in theorem II.2, $\text{var}\left(\hat{\boldsymbol{\theta}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2$ and so to determine the variance of our estimates we need to know σ^2. Generally, it is unknown and has to be estimated. Now the definition of the variance is $\text{var}[Y] = E\left[\left(Y - E[Y]\right)^2\right]$. The logical estimator is one that parallels this definition as much as possible: put in our estimate for $E[Y]$ and replace the first expectation operator by the mean.

Definition II.6: The **fitted values** are the estimated expected values for each observation. They are obtained by substituting the *x*-values for each observation into the estimated equation. They are given by $\mathbf{X}\hat{\boldsymbol{\theta}}$. The **residuals** are the deviations of the observed values of the response variable from the fitted values. They are denoted by $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}$ and are the estimates of $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\theta}$. ■

In simple linear regression, the fitted values are the points on the line corresponding to observed *x*-values and the residuals are the vertical difference between an observed point and the line.

The logical estimator for σ^2 is:

$$\hat{\sigma}_n^2 = \frac{\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}}\right)}{n} = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n}$$

and our estimate would be $\mathbf{e}'\mathbf{e}/n$.

Notice that the numerator of this expression measures the spread of observed *Y*-values from the fitted equation (line). We would expect this to be random variation.

Again we ask the question: is our estimator unbiased?

Theorem II.6: Let $\hat{\sigma}_n^2 = \left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}}\right)\Big/n$ with $\mathbf{Y}$ the $n\times 1$ random vector of sample random variables, $\mathbf{X}$ is an $n\times q$ matrix of full rank, $\hat{\boldsymbol{\theta}}$ is a $q\times 1$ vector of estimators for $\boldsymbol{\theta}$ Then $E\left[\hat{\sigma}_n^2\right] = \frac{n-q}{n}\sigma^2$ so that an unbiased estimator for σ^2 is

$$\hat{\sigma}^2_{n-q} = \frac{\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}}\right)'\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}}\right)}{n-q}$$

Proof: The proof of this theorem is postponed. ■

II.D Maximum likelihood estimation of the parameters

In previous sections we used least squares estimation to obtain estimators of our parameters. It was mentioned that this is not the only method of estimation. Here we describe an alternative method of estimation, maximum likelihood estimation. It is based on obtaining the values of the parameters that make the data most ’likely’. Compare this with least squares estimation in which the sum of squares of the differences between the data and the estimated expected values or fitted values is minimized.

Maximum likelihood estimation involves determining the likelihood function for the data and it is this function that is maximized.

Definition II.7: The **likelihood function** is the joint distribution function, $f(\mathbf{y};\boldsymbol{\xi})$, of n random variables $Y_1, Y_2, \ldots, Y_n$ evaluated at $\mathbf{y}=(y_1, y_2, \ldots, y_n)$. For fixed **y** the likelihood function is a function of the k parameters, $\boldsymbol{\xi}=(\xi_1,\xi_2,\ldots,\xi_k)$, and is denoted by $L(\boldsymbol{\xi};\mathbf{y})$. If $Y_1, Y_2, \ldots, Y_n$ represents a random sample from $f(\mathbf{y};\boldsymbol{\xi})$ then

$$L(\boldsymbol{\xi};\mathbf{y}) = f(y_1;\boldsymbol{\xi})f(y_2;\boldsymbol{\xi})\ldots f(y_n;\boldsymbol{\xi}).$$

■

Note that the notation $f(\mathbf{y};\boldsymbol{\xi})$ indicates that **y** is the vector of variables for the distribution function and that the values of ξ are fixed for the population under consideration. This is read ’the function f of **y** given ξ’. The likelihood function reverses the roles in that the variables are ξ and **y** is considered fixed so we have the likelihood of ξ given **y**.

Definition II.8: The **parameter space** for a set of parameters ξ from a probability distribution function is the set of all possible values of the parameters and is denoted by Ω. ■

For a single parameter from a continuous probability density function, the parameter space is some subset of the real numbers, R; For q parameters, it is some subset of all q-tuples of real numbers, R^q. For example, the parameter space for the parameter ψ from the normal distribution is R whereas the parameter space for the parameter σ^2 is the set of positive reals, P. The parameter space for $\boldsymbol{\xi}=\left(\psi,\sigma^2\right)$ is the set of pairs where the first element of the pair is any real number and the second element is any positive real; that is, it is the 'half' of R^2.

Definition II.9: Let $L(\boldsymbol{\xi};\mathbf{y}) = f(\mathbf{y};\boldsymbol{\xi})$ be the likelihood function for $Y_1, Y_2, \ldots, Y_n$. For a given set of observations, **y**, the value $\tilde{\boldsymbol{\xi}}$ that maximizes $L(\boldsymbol{\xi};\mathbf{y})$ is called the **maximum likelihood estimate** (MLE) of ξ. An expression for this estimate as a function of **y** can be derived. The **maximum likelihood estimators** are defined to be the same function as the estimate, with **Y** substituted for **y**. ■

The procedure for obtaining the maximum likelihood estimators is:

1. Write the expression for the distribution function of a single observation, y_i.
2. Write the likelihood function $L(\boldsymbol{\xi};\mathbf{y})$ as the product of the *n* distribution functions for the *n* observations.
3. Express $L(\boldsymbol{\xi};\mathbf{y})$ in terms of matrices.
4. Find $\ell = \ln\left[L(\boldsymbol{\xi};\mathbf{y})\right]$.
5. Maximize ℓ with respect to ξ to derive the maximum likelihood estimates.
6. Obtain the maximum likelihood estimators.

The quantity ℓ is referred to as the log likelihood. Maximizing ℓ is equivalent to maximizing $L(\boldsymbol{\xi};\mathbf{y})$ since *ln* is a monotonic function. It will turn out that the expression for ℓ is often simpler than that for $L(\boldsymbol{\xi};\mathbf{y})$.

The linear model for a response variable related to *p* explanatory variables is of the form:

$$E[Y_i] = \psi_i = \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_q x_{iq} = \sum_{j=1}^{q} \theta_j x_{ij} \text{ and } \operatorname{var}[Y_i] = \sigma^2 .$$

where $x_{i1} = 1$ and $q = p+1$.

The expectation model is of a slightly different form to that presented previously in that the intercept corresponds to θ_1 with $x_{i1} = 1$ rather than being included as θ_0 with no *x* term.

In terms of matrices, this model is

$$E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\theta} \text{ and } \operatorname{var}[\mathbf{Y}] = \mathbf{V}_{\mathbf{Y}} = \sigma^2 \mathbf{I}_n .$$

We have not previously specified a probability distribution function to be used as part of the model. However, maximum likelihood estimation requires that we do specify this function (least squares estimation does not). A distribution function commonly used for continuous variables is the normal distribution.

We now derive the maximum likelihood estimators of $\boldsymbol{\theta}$ and σ^2.

a) Maximum likelihood estimation of θ

Distribution function of y_i

The distribution function for a single observation from a normal distribution is:

$$\mathrm{f}\left(y_i;\boldsymbol{\theta}\right)=\frac{1}{\sqrt{\left(2\pi\sigma^2\right)}}\exp\left\{\frac{-\left(y_i-\psi_i\right)^2}{2\sigma^2}\right\}=\frac{1}{\sqrt{\left(2\pi\sigma^2\right)}}\exp\left\{\frac{-\left(y_i-\sum_{j=1}^{p}x_{ij}\theta_j\right)^2}{2\sigma^2}\right\}$$

The likelihood function $L\left(\boldsymbol{\theta};\mathbf{y}\right)$

$$\begin{aligned}L\left(\boldsymbol{\theta};\mathbf{y}\right)&=\prod_{i=1}^{n}\mathrm{f}\left(y_i;\boldsymbol{\theta}\right)\\&=\prod_{i=1}^{n}\left[\frac{1}{\sqrt{\left(2\pi\sigma^2\right)}}\exp\left\{\frac{-\left(y_i-\psi_i\right)^2}{2\sigma^2}\right\}\right]\\&=\prod_{i=1}^{n}\frac{1}{\sqrt{\left(2\pi\sigma^2\right)}}\prod_{i=1}^{n}\left[\exp\left\{\frac{-\left(y_i-\psi_i\right)^2}{2\sigma^2}\right\}\right]\\&=\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{\sum_{i=1}^{n}\frac{-\left(y_i-\psi_i\right)^2}{2\sigma^2}\right\}\\&=\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\theta_j\right)^2\right\}\end{aligned}$$

$L\left(\boldsymbol{\varepsilon};\mathbf{y}\right)$ in terms of matrices

To put $L\left(\boldsymbol{\varepsilon};\mathbf{y}\right)$ in terms of matrices we must re-express $\sum_{i=1}^{n}\left(y_i-\psi_i\right)^2=\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\theta_j\right)^2$ in terms of matrices. Recall that a sum of squares can be written as the product of the transposed column vector with the column vector itself. Hence,

$$\sum_{i=1}^{n}(y_i-\psi_i)^2=(\mathbf{y}-\boldsymbol{\psi})'(\mathbf{y}-\boldsymbol{\psi})=(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})$$

and so

$$\begin{aligned}L(\boldsymbol{\varepsilon};\mathbf{y})&=\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\theta_j\right)^2\right\}\\&=\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{\frac{-1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right\}\end{aligned}$$

Find $\ell=\ln\left[L(\boldsymbol{\varepsilon};\mathbf{y})\right]$

$$\begin{aligned}\ell&=\ln\left[L(\boldsymbol{\varepsilon};\mathbf{y})\right]\\&=\ln\left[\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{\frac{-1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right\}\right]\\&=-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\end{aligned}$$

Maximize ℓ with respect to $\boldsymbol{\theta}$

To maximize ℓ with respect to $\boldsymbol{\theta}$ we need to differentiate the log likelihood and set it equal to zero.

$$\begin{aligned}\frac{\partial\ell}{\partial\boldsymbol{\theta}}&=\frac{\partial\left\{-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right\}}{\partial\boldsymbol{\theta}}\\&=-\frac{1}{2\sigma^2}\frac{\partial\left\{(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right\}}{\partial\boldsymbol{\theta}}\\&=\mathbf{0}\end{aligned}$$

That is, the maximum likelihood estimates will be the solution of

$$\frac{\partial\left\{(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right\}}{\partial\boldsymbol{\theta}}=\mathbf{0}$$

But this is exactly the same expression, except it involves **y** instead of **Y**, that had to be minimized in Theorem II.1 to obtain the least squares estimators.

So the maximum likelihood estimates are $\tilde{\boldsymbol{\theta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Obtain the maximum likelihood estimators

Finally the maximum likelihood estimators are obtained by substituting **Y** for **y** in the expression for the estimates and so are $\tilde{\boldsymbol{\theta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$.

We formalize this result as a theorem.

Theorem II.7: Let **Y** be a normally distributed random vector representing a random sample with $E\left[\mathbf{Y}\right] = \mathbf{X}\boldsymbol{\theta}$ and $\text{var}\left[\mathbf{Y}\right] = \mathbf{V}_{\mathbf{Y}} = \sigma^2\mathbf{I}_n$ where **X** is an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ is a $q \times 1$ vector of unknown parameters and $n \geq q$. The maximum likelihood estimator for $\boldsymbol{\theta}$ is denoted by $\tilde{\boldsymbol{\theta}}$ and is given by

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

Proof: see above ■

So for linear models the maximum likelihood and least squares estimators coincide. It turns out that maximum likelihood estimators have some further desirable properties so that it is an advantage to know that our estimators are maximum likelihood estimators.

b) Maximum likelihood estimation of σ^2

Theorem II.8: Let **Y** be a normally distributed random vector representing a random sample with $E\left[\mathbf{Y}\right] = \mathbf{X}\boldsymbol{\theta}$ and $\text{var}\left[\mathbf{Y}\right] = \mathbf{V}_{\mathbf{Y}} = \sigma^2\mathbf{I}_n$ where **X** is an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ is a $q \times 1$ vector of unknown parameters and $n \geq q$. The maximum likelihood estimator for σ^2 is denoted by $\tilde{\sigma}^2$ and is given by

$$\tilde{\sigma}_n^2 = \frac{\left(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\theta}}\right)'\left(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\theta}}\right)}{n} = \frac{\tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}}{n}$$

Proof: The proof is left as an exercise for you ■

II.E ANOVA method of hypothesis testing in regression

It is often useful to be able perform hypothesis tests about the composition of the model. For example, to test whether the model is useful at all, we could perform an hypothesis test for H$_0$: $\boldsymbol{\theta} = \mathbf{0}$ against the alternative hypothesis H$_a$: $\boldsymbol{\theta} \neq \mathbf{0}$. In the analysis of variance (ANOVA) method a sum of squares is divided into components that that can be attributed to important sources of variation. These components are then compared to test the hypothesis.

a) The analysis of variance

For the hypothesis test just mentioned, the Total sum of squares is divided into the Regression and Residual sums of squares.

Definition II.10: Let $\hat{\boldsymbol{\theta}}$ be the least squares estimates of $\boldsymbol{\theta}$ so that $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ where $\mathbf{y}$ is an $n \times 1$ vector of observations from a random sample, $\mathbf{X}$ is an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ is a $q \times 1$ vector of unknown parameters and $n \geq q$. Then define $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{R}_\mathbf{X} = \mathbf{I} - \mathbf{P}_\mathbf{X}$ so that the fitted values are $\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{P}_\mathbf{X}\mathbf{y}$ and the residuals are $\mathbf{e} = \mathbf{R}_\mathbf{X}\mathbf{y}$. ■

It is easy to see that $\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{P}_\mathbf{X}\mathbf{y}$. Note that in definition II.6, the residuals were defined to be $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}$ and so $\mathbf{e} = \mathbf{y} - \mathbf{P}_\mathbf{X}\mathbf{y} = (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{y} = \mathbf{R}_\mathbf{X}\mathbf{y}$ as defined.

Theorem II.9: Let $\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ be the least squares estimates of $\boldsymbol{\theta}$ where $\mathbf{y}$ is an $n \times 1$ vector of observations from a random sample, $\mathbf{X}$ is an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ is a $q \times 1$ vector of unknown parameters and $n \geq q$. The Total sum of squares of the observations, $\mathbf{y}'\mathbf{y}$, can be expressed as the sum of two sums of squares as follows:

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} + \mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y}.$$

where $\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y}$ is the sum of squares of the fitted values and $\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y}$ is the sum of squares of the residuals.

Proof: First we prove the sums of squares identity:

$$\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} + \mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y} = \mathbf{y}'(\mathbf{P}_\mathbf{X} + \mathbf{R}_\mathbf{X})\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y} = \mathbf{y}'\mathbf{y}.$$

Next, we have to show that $\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} = \left(\mathbf{X}\hat{\boldsymbol{\theta}}\right)'\mathbf{X}\hat{\boldsymbol{\theta}}$. We note that it can be proved that $\mathbf{P}_\mathbf{X}$ is symmetric and idempotent in that $\mathbf{P}_\mathbf{X}' = \mathbf{P}_\mathbf{X}$ and $\mathbf{P}_\mathbf{X}^2 = \mathbf{P}_\mathbf{X}$. Consequently,

$$\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} = \mathbf{y}'\mathbf{P}_\mathbf{X}'\mathbf{P}_\mathbf{X}\mathbf{y} = (\mathbf{P}_\mathbf{X}\mathbf{y})'\mathbf{P}_\mathbf{X}\mathbf{y} = \left(\mathbf{X}\hat{\boldsymbol{\theta}}\right)'\mathbf{X}\hat{\boldsymbol{\theta}}$$

Finally, we have to prove that $\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y} = \mathbf{e}'\mathbf{e}$ and we note that it can be proved that $\mathbf{R}_\mathbf{X}$ is symmetric and idempotent in that $\mathbf{R}_\mathbf{X}' = \mathbf{R}_\mathbf{X}$ and $\mathbf{R}_\mathbf{X}^2 = \mathbf{R}_\mathbf{X}$. Consequently,

$$\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y} = (\mathbf{R}_\mathbf{X}\mathbf{y})'\mathbf{R}_\mathbf{X}\mathbf{y} = \mathbf{e}'\mathbf{e}$$

as required. ■

So $\mathbf{y}'\mathbf{y}$ can be partitioned into $\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y} = \hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}}$ and $\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y} = \mathbf{e}'\mathbf{e}$. The latter two terms are called the Regression and Residual sums of squares, respectively. Generally, the Regression sum of squares is computed directly from $\hat{\boldsymbol{\theta}}$ and the Regression sum of squares is computed as the difference of the other two sums of squares. In section II.C, we noted that the residuals are the vertical deviations of the observed data from the fitted surface (line). We would expect these to be random and so the Residual sum of squares would reflect the random variation around the fitted line. The Regression sum of squares is just the sum of squares of the fitted values. It reflects the variation along the fitted surface (line). We would hope that the Regression sum of squares is large relative to the Residual sum of squares.

As well as the sums of squares we also require the degrees of freedom of the sums of squares. Note that the three sums of squares are all quadratic forms in **y**, the matrices of the quadratic forms being $\mathbf{I}_\mathrm{n}$, $\mathbf{P}_\mathbf{X}$ and $\mathbf{R}_\mathbf{X}$ and that they are the sums of squares of $\mathbf{I}_n\mathbf{y}$, $\mathbf{P}_\mathbf{X}\mathbf{y}$ and $\mathbf{R}_\mathbf{X}\mathbf{y}$. If we define the degrees of freedom of a sum of squares to be the number of independent quantities that are squared and summed, then it can be shown that the degrees of freedom is equal to the rank of the matrix of the quadratic form. Hence to get the degrees of freedom we require the ranks.

Theorem II.10: Let $\mathbf{y}'\mathbf{y}$ be the Total sums of squares, $\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y}$ be the Regression sums of squares and $\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y}$ be the Residual sums of squares where **y** is an $n\times1$ vector of observed values from a random sample and $\mathbf{P}_\mathbf{X}$ and $\mathbf{R}_\mathbf{X}$ are as given in definition II.10. Then the degrees of freedom of $\mathbf{y}'\mathbf{y}$ is *n*, of $\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y}$ is *q* and of $\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y}$ is *n*–*q*.

Proof: The degrees of freedom of $\mathbf{y}'\mathbf{y}$ is given by $rank(\mathbf{I}_n)$ which is clearly *n*.

The degrees of freedom of $\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y}$ is given by $rank(\mathbf{P}_\mathbf{X})$. Now, for **B** idempotent, $rank(\mathbf{B}) = trace(\mathbf{B})$ as:

- the trace of a matrix is equal to the sum of its eigenvalues;
- the rank of a matrix is equal to the number of nonzero eigenvalues; and
- the eigenvalues of an idempotent can be proved to be either 1 or zero.

Hence,

$$\begin{aligned} rank(\mathbf{P}_\mathbf{X}) &= trace(\mathbf{P}_\mathbf{X}) \\ &= trace\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\ &= trace\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right) \\ &= trace(\mathbf{I}_q) \\ &= q \end{aligned}$$

Note for first step use the fact that the trace of matrix products are cyclically commutative; that is,

$$trace(\mathbf{ABC}) = trace(\mathbf{BCA}) = trace(\mathbf{CAB}).$$

The degrees of freedom of $\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y}$ is given by

$$\begin{aligned} rank &= trace(\mathbf{R}_\mathbf{X}) \\ &= trace\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\ &= trace(\mathbf{I}_n) - trace\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right) \\ &= n - q \end{aligned}$$

Note for first step use $trace(\mathbf{A}+\mathbf{C}) = trace(\mathbf{A}) + trace(\mathbf{C})$.

■

It is convenient to summarize our computations in a table, the analysis of variance table.

Source	**DF**	**SSq**	**MSq**	**F**
Regression (Linear)	q	$\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y}$	$\dfrac{\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y}}{q}\ (= MS_L)$	$\dfrac{MS_L}{MS_R}$
Residual	n-q	$\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y}$	$\dfrac{\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y}}{n-q}\ (= MS_R)$	
Total	n	$\mathbf{y}'\mathbf{y}$		

Example II.1 House price (continued)

For this example, involving the explanatory variable Age and Area, we have seen that

$$\mathbf{y} = \begin{bmatrix} 50 \\ 40 \\ 52 \\ 47 \\ 65 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 10 & 2 \\ 1 & 20 & 3 \end{bmatrix} \text{ and } \hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} 33.06 \\ -0.189 \\ 10.178 \end{bmatrix}$$

From these we obtain the residuals and fitted values:

	Observations $(=\mathbf{y})$	Fitted values $\left(=\mathbf{X}\hat{\boldsymbol{\theta}}=\mathbf{P}_{\mathbf{X}}\mathbf{y}\right)$	Residuals $\left(=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\theta}}=\mathbf{R}_{\mathbf{X}}\mathbf{y}\right)$
	50	43.589	6.411
	40	42.833	-2.833
	52	53.551	-1.551
	47	52.606	-5.606
	65	61.434	3.566
SSq	13238	13143.904	95.676

Note that the sum of the last two sums of squares is 13239.580 which is the same as the first sum of squares, except for rounding error. The analysis of variance table for the example is:

Source	**DF**	**SSq**	**MSq**	**F**
Regression	3	13143.904	4381.3013	93.12
Residual	2	94.096	47.0480	
Total	5	13238.000		

b) Expected values of the mean squares

The justification for this analysis can be made more substantial by considering the expected values of the mean squares: $E[MS_L]$ and $E[MS_R]$. The expected values of the mean squares are just the average or mean value of the mean squares under sampling from a population in which the variables *X* and *Y* are related as specified by the linear regression model To derive the expected values, we note that the general form of the two mean squares is a quadratic form divided by a degrees of freedom. So we first investigate the expectation of a quadratic form.

Theorem II.11: Let **Y** be an $n\times 1$ vector of random variables with $E[\mathbf{Y}]=\boldsymbol{\psi}$ and $\text{var}[\mathbf{Y}]=\mathbf{V}$, where $\boldsymbol{\psi}$ is a $n\times 1$ vector of expected values and **V** is an $n\times n$ matrix. Let **B** an $n\times n$ matrix of real numbers. Then

$$E[\mathbf{Y}'\mathbf{B}\mathbf{Y}]=trace(\mathbf{B}\mathbf{V})+\boldsymbol{\psi}'\mathbf{B}\boldsymbol{\psi}.$$

Proof: Firstly note that, as $\mathbf{Y}'\mathbf{B}\mathbf{Y}$ is a scalar, $\mathbf{Y}'\mathbf{B}\mathbf{Y}=trace(\mathbf{Y}'\mathbf{B}\mathbf{Y})$ and recall that the trace of matrix products are cyclically commutative.

Thus, $E[\mathbf{Y}'\mathbf{B}\mathbf{Y}]=E\left[trace(\mathbf{Y}'\mathbf{B}\mathbf{Y})\right]=E\left[trace(\mathbf{B}\mathbf{Y}\mathbf{Y}')\right]$

Now, it can be shown that $E\left[trace(\mathbf{A})\right]=trace\left(E[\mathbf{A}]\right)$; this result is derived from the fact that the expectation of the sum of a set of quantities is equal to the sum of their expectations as the trace is just the sum of the diagonal elements of **A**. Also, theorem II.3 states that $E[\mathbf{A}\mathbf{Y}]=\mathbf{A}E[\mathbf{Y}]$.

Hence, $E\left[trace\left(\mathbf{BYY}'\right)\right] = trace\left(E\left[\mathbf{BYY}'\right]\right) = trace\left(\mathbf{B}E\left[\mathbf{YY}'\right]\right)$.

By definition $\mathbf{V} = E\left[\left(\mathbf{Y}-\boldsymbol{\psi}\right)\left(\mathbf{Y}-\boldsymbol{\psi}\right)'\right]$ so that $\mathbf{V} = E\left[\mathbf{YY}' - \mathbf{Y}\boldsymbol{\psi}' - \boldsymbol{\psi}\mathbf{Y}' + \boldsymbol{\psi}\boldsymbol{\psi}'\right]$.

Now, $E\left[\boldsymbol{\psi}\right] = \boldsymbol{\psi}$ as the elements of $\boldsymbol{\psi}$ are population quantities and their average values in the population are the values to which they are equal. That is, they are constants with respect to expectation. Thus

$$E\left[\mathbf{Y}\boldsymbol{\psi}'\right] = E\left[\boldsymbol{\psi}\mathbf{Y}'\right] = \boldsymbol{\psi}\boldsymbol{\psi}'$$

Hence, $\mathbf{V} = E\left[\mathbf{YY}'\right] - \boldsymbol{\psi}\boldsymbol{\psi}'$ so that $E\left[\mathbf{YY}'\right] = \mathbf{V} + \boldsymbol{\psi}\boldsymbol{\psi}'$.

This leads to

$$\begin{aligned} E\left[\mathbf{Y}'\mathbf{BY}\right] &= trace\left(\mathbf{B}\left(\mathbf{V}+\boldsymbol{\psi}\boldsymbol{\psi}'\right)\right) \\ &= trace\left(\mathbf{BV}\right) + trace\left(\mathbf{B}\boldsymbol{\psi}\boldsymbol{\psi}'\right) \\ &= trace\left(\mathbf{BV}\right) + trace\left(\boldsymbol{\psi}'\mathbf{B}\boldsymbol{\psi}\right) \\ &= trace\left(\mathbf{BV}\right) + \boldsymbol{\psi}'\mathbf{B}\boldsymbol{\psi} \end{aligned}$$

■

Theorem II.12: Let $MS_L = \mathbf{Y}'\mathbf{P_X}\mathbf{Y}/q$ be the estimator for the Regression sums of squares where $\mathbf{P_X} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, $E\left[\mathbf{Y}\right] = \mathbf{X}\boldsymbol{\theta}$ and $\text{var}\left[\mathbf{Y}\right] = \sigma^2\mathbf{I}_n$ with $\mathbf{X}$ an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ a $q \times 1$ vector of unknown parameters and $n \geq q$. Then

$$E\left[MS_L\right] = E\left[\mathbf{Y}'\mathbf{P_X}\mathbf{Y}/q\right] = \sigma^2 + \left(1/q\right)\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}$$

Proof:

$$\begin{aligned} E\left[MS_L\right] &= E\left[\mathbf{Y}'\mathbf{P_X}\mathbf{Y}/q\right] \\ &= E\left[\mathbf{Y}'\mathbf{P_X}\mathbf{Y}\right]/q \\ &= \left\{trace\left(\mathbf{P_X}\sigma^2\mathbf{I}_n\right) + \boldsymbol{\theta}'\mathbf{X}'\mathbf{P_X}\mathbf{X}\boldsymbol{\theta}\right\}/q \\ &= \left\{\sigma^2 trace\left(\mathbf{P_X}\right) + \boldsymbol{\theta}'\mathbf{X}'\mathbf{P_X}\mathbf{X}\boldsymbol{\theta}\right\}/q \end{aligned}$$

Now $\boldsymbol{\theta}'\mathbf{X}'\mathbf{P_X}\mathbf{X}\boldsymbol{\theta} = \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}$ and it was proved in theorem II.10 that $trace\left(\mathbf{P_X}\right) = q$.

Hence,

$$\begin{aligned} E[MS_L] &= \left\{\sigma^2 trace(\mathbf{P_X}) + \boldsymbol{\theta}'\mathbf{X}'\mathbf{P_X}\mathbf{X}\boldsymbol{\theta}\right\}\Big/q \\ &= \left\{q\sigma^2 + \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}\right\}\Big/q \\ &= \sigma^2 + (1/q)\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} \end{aligned}$$

■

Theorem II.13: Let $MS_R = \mathbf{Y}'\mathbf{R_X}\mathbf{Y}/(n-q)$ be the estimator for the Residual sums of squares where $\mathbf{R_X} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\theta}$ and $\mathrm{var}[\mathbf{Y}] = \sigma^2\mathbf{I}_n$ with $\mathbf{X}$ an $n\times q$ matrix of full rank, $\boldsymbol{\theta}$ a $q\times 1$ vector of unknown parameters and $n \geq q$. Then

$$E[MS_R] = E\left[\mathbf{Y}'\mathbf{R_X}\mathbf{Y}/(n-q)\right] = \sigma^2$$

Proof:

$$\begin{aligned} E[MS_R] &= E[\mathbf{Y}'\mathbf{R_X}\mathbf{Y}]/(n-q) \\ &= \left\{trace\left(\mathbf{R_X}\sigma^2\mathbf{I}_n\right) + \boldsymbol{\theta}'\mathbf{X}'\mathbf{R_X}\mathbf{X}\boldsymbol{\theta}\right\}\Big/(n-q) \\ &= \left\{\sigma^2 trace(\mathbf{R_X}) + \boldsymbol{\theta}'\mathbf{X}'\mathbf{R_X}\mathbf{X}\boldsymbol{\theta}\right\}\Big/(n-q) \end{aligned}$$

Now and it was proved in theorem II.10 that $trace(\mathbf{R_X}) = n-q$.

Also,

$$\begin{aligned} \boldsymbol{\theta}'\mathbf{X}'\mathbf{R_X}\mathbf{X}\boldsymbol{\theta} &= \boldsymbol{\theta}'\mathbf{X}'\left\{\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right\}\mathbf{X}\boldsymbol{\theta} \\ &= \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\theta} \\ &= \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} \\ &= 0 \end{aligned}$$

Hence,

$$\begin{aligned} E[MS_R] &= \left\{\sigma^2 trace(\mathbf{R_X}) + \boldsymbol{\theta}'\mathbf{X}'\mathbf{R_X}\mathbf{X}\boldsymbol{\theta}\right\}\Big/(n-q) \\ &= \left\{(n-q)\sigma^2\right\}\Big/(n-q) \\ &= \sigma^2 \end{aligned}$$

■

Note that MS$_R$ is nothing more than $\hat{\sigma}^2_{n-q} = \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\theta}}\right)\Big/(n-q)$ from theorem II.6. This theorem claimed that it was an unbiased estimator of σ^2 and theorem II.13 proves that it is.

We now add the expected mean squares to the analysis of variance table.

Source	DF	MSq	E[MSq]	F
Regression (Linear)	q	$\dfrac{\mathbf{y}'\mathbf{P}_\mathbf{X}\mathbf{y}}{q}$	$\sigma^2 + (1/q)\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}$	$\dfrac{MS_L}{MS_R}$
Residual	n-q	$\dfrac{\mathbf{y}'\mathbf{R}_\mathbf{X}\mathbf{y}}{n-q}$	σ^2	
Total	n	$\mathbf{y}'\mathbf{y}$		

Now we note that the difference between the two expected mean squares is $\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}$ which is zero when $\boldsymbol{\theta} = \mathbf{0}$ as hypothesized under the null hypothesis. In this case we would expect our test statistic to be about 1. But what about when the null hypothesis is not true? That is, $\boldsymbol{\theta} \neq \mathbf{0}$? Note that $\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} \geq 0$ because it is the sum of squares of the elements of $\mathbf{X}\boldsymbol{\theta}$. However, is it true that for $\boldsymbol{\theta} \neq \mathbf{0}$, $\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} > 0$? The next theorem proves that it is. And so we expect a value greater than one for the ratio of the mean squares when the null hypothesis is not true.

Theorem II.14: Let $\mathbf{X}$ be an $n\times q$ matrix of full rank, $\boldsymbol{\theta}$ a $q\times 1$ vector of unknown parameters and $n \geq q$. Then, for $\boldsymbol{\theta} \neq \mathbf{0}$, $\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} > 0$.

Proof: Since $\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = (\mathbf{X}\boldsymbol{\theta})'\,\mathbf{X}\boldsymbol{\theta}$ is a sums of squares, it must be nonnegative.

We need to prove that if $\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = 0$, then $\boldsymbol{\theta} = \mathbf{0}$.

To do this note that if $\boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = 0$, then $\mathbf{X}\boldsymbol{\theta} = \mathbf{0}$ and $\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = \mathbf{X}'\mathbf{0} = \mathbf{0}$.

Since $\mathbf{X}'\mathbf{X}$ is nonsingular, $(\mathbf{X}'\mathbf{X})^{-1}$ exists and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{0}$ so that $\boldsymbol{\theta} = \mathbf{0}$.

■

c) Distribution of the test statistic

In order to assess whether or not our null hypothesis is likely to be true, we want to know if the value of our test statistic is unlikely when the null hypothesis is true. To do this we need to establish the sampling distribution of the test statistic when the null hypothesis is true. The sampling distribution of the test statistic is the distribution that would be obtained under repeated sampling of the population. To construct this sampling distribution you would have to find a population in which you knew that the null hypothesis was true, take many samples from the population, for each sample compute the test statistic and construct the distribution from the computed test statistics. Not very practical! However, the following theorems will help.

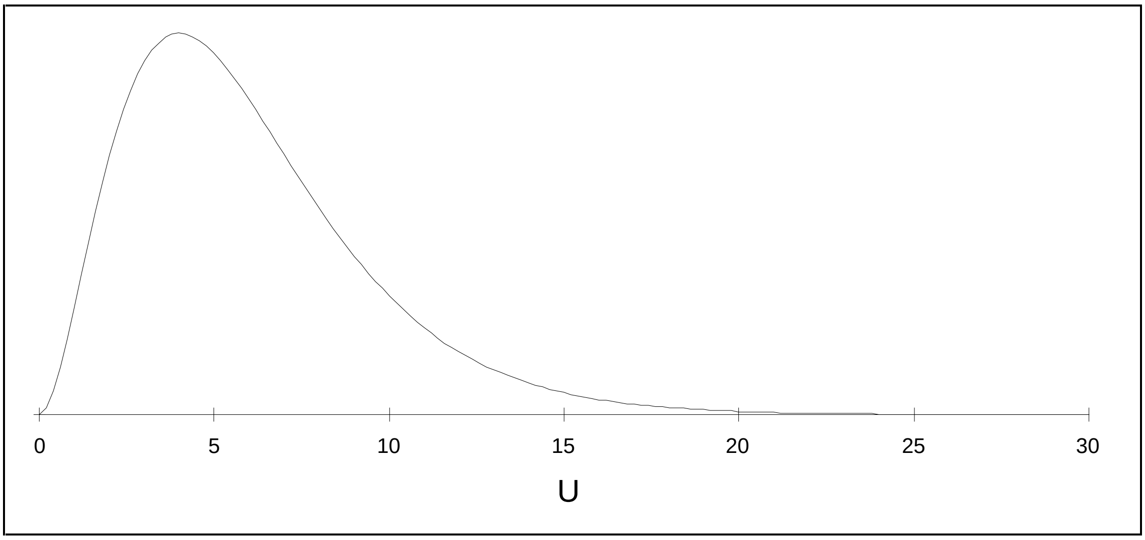
Theorem II.15: Let $\mathbf{A}$ be an $n\times n$ symmetric matrix and $\mathbf{Y}$ be an $n\times 1$ normally distributed random vector with $E[\mathbf{AY}] = \mathbf{0}$ and $\mathrm{var}[\mathbf{Y}] = \sigma^2\mathbf{I}_n$. Then $(1/\sigma^2)\mathbf{y}'\mathbf{A}\mathbf{y}$ follows a chi-squared distribution with r degrees of freedom if and only if $\mathbf{A}$ is idempotent of rank r.

Proof: not given ■

The chi-square probability distribution function for the random variable *U* is:

$$\chi_n^2(u) = \left(\frac{1}{\Gamma(n/2)2^{n/2}}\right)u^{(n-2)/2}e^{-u/2}, \qquad 0 < u < \infty$$

Probability distribution function for χ_6^2



Theorem II.16: Let **Y** be an $n\times 1$ normally distributed random vector with $E[\mathbf{Y}] = \mathbf{X\theta}$ and $\text{var}[\mathbf{Y}] = \sigma^2\mathbf{I}_n$. Let $\mathbf{Y'A}_1\mathbf{Y}, \mathbf{Y'A}_2\mathbf{Y}, \ldots, \mathbf{Y'A}_m\mathbf{Y}$ be a collection of *m* quadratic forms where, for each $i = 1, 2, \ldots, m$, $\mathbf{A}_1$ is symmetric, of rank r_i and $E[\mathbf{A}_i\mathbf{Y}] = \mathbf{0}$. If any two of the following three statements are true, then for each *i*, $\left(1/\sigma^2\right)\mathbf{Y'A}_i\mathbf{Y}$ follows a chi-squared distribution with r_i degrees of freedom. Furthermore, $\mathbf{Y'A}_i\mathbf{Y}$ are independent for $i \neq j$ and $\sum_{i=1}^m r_i = r$ where *r* denotes the rank of $\sum_{i=1}^m \mathbf{A}_i$.

1. All $\mathbf{A}_\text{i}$ are idempotent
2. $\sum_{i=1}^m \mathbf{A}_i$ is idempotent
3. $\mathbf{A}_i\mathbf{A}_j = \mathbf{0}, \quad i \neq j$

Proof: not given ■

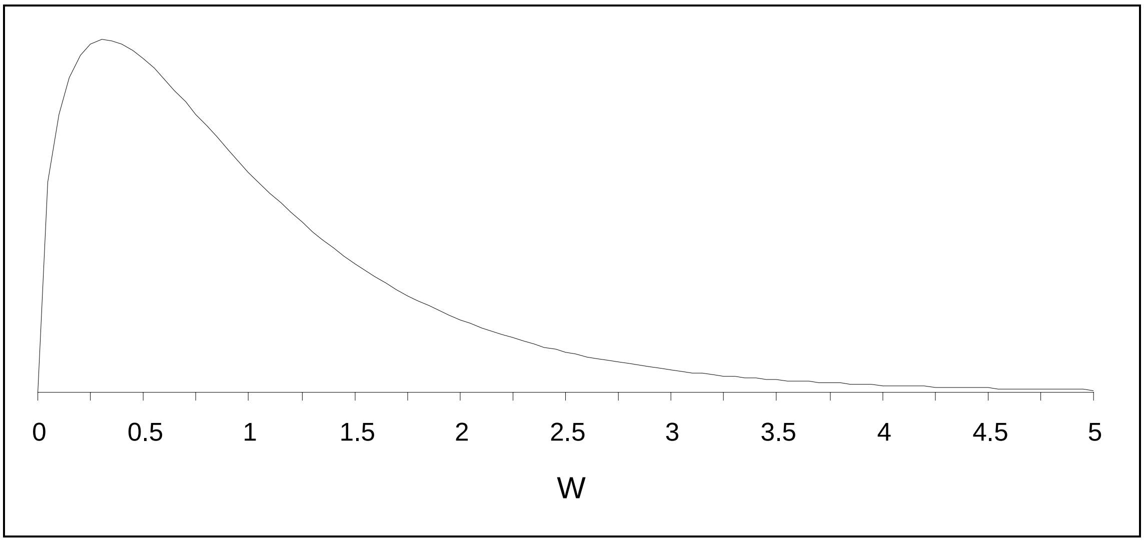
Theorem II.17: Let U_1 and U_2 be two random variables distributed as chi-squares with r_1 and r_2 degrees of freedom. Then, provided U_1 and U_2, are independent, the random variable $W = \dfrac{U_1/r_1}{U_2/r_2}$ is distributed as Snedecor's F with r_1 and r_2 degrees of freedom.

Proof: not given ■

The F probability distribution function for the random variable W is:

$$F_{r_1,r_2}(w) = \left(\frac{\Gamma\left(\frac{r_1+r_2}{2}\right)\left(\frac{r_1}{r_2}\right)^{r_1/2}}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)} \right) w^{(r_1-2)/2}\left(1+\left(\frac{r_1}{r_2}\right)w\right)^{-(r_1+r_2)/2}, \quad 0 < w < \infty$$

Probability distribution function for $\mathbf{F_{3,46}}$



Theorem II.18: Let $MS_L = \mathbf{Y'P_XY}/q$ and $MS_R = \mathbf{Y'R_XY}/(n-q)$ be the estimators for the Regression and Residual sums of squares, respectively, where $\mathbf{P_X} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$, $\mathbf{R_X} = \mathbf{I}_n - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$, $\mathbf{Y}$ be a normally distributed random vector, $E[\mathbf{Y}] = \mathbf{0}$ and $\text{var}[\mathbf{Y}] = \sigma^2\mathbf{I}_n$ with $\mathbf{X}$ an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ a $q \times 1$ vector of unknown parameters and $n \geq q$. Then, the ratio of these two mean squares, given by

$$F_{q,\,n-q} = \frac{\mathbf{Y'P_XY}/q}{\mathbf{Y'R_XY}/(n-q)}$$

is distributed as a Snedecor's F with q and $(n-q)$ degrees of freedom.

Proof: Since $\mathbf{I}_n = \mathbf{P_X} + \mathbf{R_X}$, the sum of the two symmetric idempotents is itself clearly idempotent. Also, $E[\mathbf{P_XY}] = E[\mathbf{R_XY}] = \mathbf{0}$ because $E[\mathbf{Y}] = \mathbf{0}$. Hence, theorem II.16

applies and the first two conditions of the three listed are met so that $\left(1/\sigma^2\right)\mathbf{Y'P_XY}$ and $\left(1/\sigma^2\right)\mathbf{Y'R_XY}$ are distributed independently as chi-squares with q and n-q degrees of freedom, respectively. Therefore, by theorem II.17, the ratio of these two quantities, each divided by their degrees of freedom, is distributed as an F with q and $(n-q)$ degrees of freedom. That is,

$$F_{q,\, n-q} = \frac{\left(1/\sigma^2\right)\mathbf{Y'P_XY}/q}{\left(1/\sigma^2\right)\mathbf{Y'R_XY}/(n-q)} = \frac{\mathbf{Y'P_XY}/q}{\mathbf{Y'R_XY}/(n-q)}$$

is distributed as a Snedecor's F with q and $(n-q)$ degrees of freedom as claimed. ■

Given this result we can now formulate a test for the hypothesis test we have been considering. In particular, we note that the null hypothesis is H_0: $\boldsymbol{\theta} = \mathbf{0}$ against the alternative hypothesis H_a: $\boldsymbol{\theta} \neq \mathbf{0}$. Consequently, under the null hypothesis, $E[\mathbf{Y}] = \mathbf{0}$ and we can use theorem II.18 to conclude that the test statistic computed using the analysis of variance table is distributed as a Snedecor's F with q and $(n-q)$ degrees of freedom. Of course, this is only true if the null hypothesis is true. So we can use this distribution to determine how unlikely is our observed value of the test statistic when the null hypothesis is true. Using a computer to do the computations, I find that the probability of a value of $F_{3,2}$= 93.12 or larger is 0.0106. We are not likely to get this value if H_o is true. We reject the null hypothesis and conclude that the evidence suggests that $\boldsymbol{\theta} \neq \mathbf{0}$.

Note that we have now performed a complete statistical inference procedure as follows:

1. *Specify the parameter of interest:* as far as the hypothesis test is concerned the **parameter** is the population value of the ratio of the mean squares, say $\Phi_{3,\infty}$.
2. *Set up a probability model for the situation:* the **probability model** is that the population distribution is described by the normal distribution with
$$E[Y] = \psi = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_q x_q = \textstyle\sum_{i=1}^{q} \theta_j x_j \text{ and } \operatorname{var}[Y] = \sigma^2 .$$
3. *Obtain a sample, identify a statistic to estimate the parameter and compute the statistic:* The estimator for the ratio of the mean squares is $F_{q,\, n-q}$; the formula is used to compute the estimate or **statistic** which for the example is $F_{3,2}$= 93.12.
4. *Derive the sampling distribution of the statistic:* We showed that the sampling distribution is Snedecor's F distribution.
5. *Make inference about the parameter:* Using the sampling distribution we found that the probability of a value of $F_{3,2}$= 93.12 or larger is 0.0106 from which we infer that the null hypothesis should be rejected.

d) Hypothesis tests on subvectors

The hypothesis tested in the last section, H_0: $\boldsymbol{\theta} = \mathbf{0}$ is rarely tenable. Most measurements are not on average zero. For this reason, it is unusual for the intercept to be zero. So it at least is usually in the model. In fact, it should always be included unless you know that for all explanatory variables equal to 0, the response variable **must** be zero **and** you are certain that the relationship is modelled by the fitted linear model even as the explanatory variables approach 0. So, in simple linear regression the relationship must be linear from the largest observed value all the way down to zero. It is rare that both of these conditions are satisfied.

Also, we in general have *p* explanatory variables and we will usually want to examine whether some subset of their coefficients is zero, rather than being restricted to all of them being zero. This will allow us to investigate which of the explanatory variables are related to the response variable. To do this we need to develop a method of testing hypotheses concerning arbitrary subsets of the parameters $\boldsymbol{\theta}$.

Consider any subset of *s* parameters chosen from $\boldsymbol{\theta}$. Without loss of generality, it can be assumed that the first *s* parameters are selected. Now partition $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_s \\ \text{—} \\ \theta_{s+1} \\ \theta_{s+2} \\ \vdots \\ \theta_q \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are $s \times 1$ and $(q-s) \times 1$ column vectors. The matrix $\mathbf{X}$ can be partitioned as $[\mathbf{X}_1 \quad \mathbf{X}_2]$ where $\mathbf{X}_1$ contains the columns of $\mathbf{X}$ corresponding to the elements of $\boldsymbol{\theta}_1$ and $\mathbf{X}_2$ are the columns of $\mathbf{X}$ corresponding to the elements of $\boldsymbol{\theta}_2$. We want to test

$$\text{H}_0\text{: } \boldsymbol{\theta}_1 = \mathbf{0} \text{ against H}_a\text{: } \boldsymbol{\theta}_1 \neq \mathbf{0}.$$

Practically speaking, we are testing the null hypothesis that the first *s* parameters are not needed to explain the variation in the response variable versus the alternative that they are needed. Mathematically, two models are being compared. Under H_0 the models is that $E[\mathbf{Y}] = \mathbf{X}_2\boldsymbol{\theta}_2$ and this model is called a **reduced model** because it involves only $(q-s)$ parameters. Under the alternative hypothesis, the model is $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\theta} = \mathbf{X}_1\boldsymbol{\theta}_1 + \mathbf{X}_2\boldsymbol{\theta}_2$ which is called the **full model** and is based on *q* parameters. In choosing between the reduced model and the full model, we will retain the reduced model only if it is demonstrated to be adequate.

We need a test statistic! The logic is straightforward. However, it will be easier to see if we use the $D(\cdot)$ and $R(\cdot)$ notation for sums of squares.

Definition II.11: The estimator of the Residual sums of squares or **deviance** after fitting the model involving parameters $\boldsymbol{\theta}$ is denoted $D(\boldsymbol{\theta})$. The estimator for the Residual sum of squares after fitting the **null model** $E[\mathbf{Y}] = \mathbf{0}$ is $\mathbf{Y}'\mathbf{Y}$. The estimator of the **reduction in sum of squares** in going from one model that involves parameters, say $\boldsymbol{\theta}_2$, to a second model that involves more parameters, say $\boldsymbol{\theta}$ made up of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, is denoted $R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) = D(\boldsymbol{\theta}_2) - D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ and is read 'the reduction in the sums of squares for including the parameters $\boldsymbol{\theta}_1$ in the model given that the parameters $\boldsymbol{\theta}_2$ are already in the model'. The estimator of the reduction in the sum of squares for adding parameters $\boldsymbol{\theta}' = [\boldsymbol{\theta}_1,\boldsymbol{\theta}_2]'$ to the null model is denoted $R(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbf{Y}'\mathbf{Y} - D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$; that is, the model parameters $\mathbf{0}$ are not included. ■

So returning to the problem of finding a test statistic for deciding between the full and reduced model, we have that the reduction in the sum of squares for the full model given the null model is $R(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbf{Y}'\mathbf{Y} - D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \mathbf{Y}'\mathbf{P}_\mathbf{X}\mathbf{Y}$ — the Regression sums of squares for the full model. Clearly, the reduction in the sum of squares for the reduced model given the null model is $R(\boldsymbol{\theta}_2) = \mathbf{Y}'\mathbf{Y} - D(\boldsymbol{\theta}_2) = \mathbf{Y}'\mathbf{P}_{\mathbf{X}_2}\mathbf{Y}$ — the Regression sums of squares for the reduced model. So how much extra variation is explained by adding the parameters $\boldsymbol{\theta}_1$ to the model that already has parameters $\boldsymbol{\theta}_2$ in the model? The obvious thing to do is to take the difference between the variation explained with all parameters in the model and the variation explained with just the parameters $\boldsymbol{\theta}_2$ in the model:

$$\begin{aligned} R(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) - R(\boldsymbol{\theta}_2) &= R(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) - R(\boldsymbol{\theta}_2) \\ &= \{\mathbf{Y}'\mathbf{Y} - D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\} - \{\mathbf{Y}'\mathbf{Y} - D(\boldsymbol{\theta}_2)\} \\ &= D(\boldsymbol{\theta}_2) - D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \\ &= R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) \end{aligned}$$

That we can get the reduction sum of squares as the difference of either Residual or Regression sums of squares from the two models. However, the difference of the Residual sums of squares is the preferred method as it is more general.

We would expect this difference to be small if $\boldsymbol{\theta}_1$ made little difference and to be large if it made a big difference. An obvious test statistic is

$$\frac{R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)/s}{D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)/(n-q)} = F_{s,(n-q)}$$

However to establish the distribution of such a test statistic we need to know if $R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)$ is quadratic form independent of the quadratic form $D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, the latter being the Residual sum of square for the full model. In order to determine this we first establish that the matrix of the quadratic form is a symmetric idempotent.

Theorem II.19: Let $R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)=\mathbf{Y}'\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{Y}$ be the reduction in sum of squares for adding the parameters $\boldsymbol{\theta}_1$ to a model that has parameters $\boldsymbol{\theta}_2$ where $\boldsymbol{\theta}_1$ is an $s\times 1$ column vector, $\boldsymbol{\theta}_2$ is an $(q-s)\times 1$ column vector, $\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}=\mathbf{P}_{\mathbf{X}}-\mathbf{P}_{\mathbf{X}_2}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$, $\mathbf{X}$ is an $n\times q$ matrix and $\mathbf{X}_2$ is an $n\times(q-s)$ matrix. Then $\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}$ is a symmetric idempotent of rank *s*.

Proof: Since $(\mathbf{A}+\mathbf{B})'=\mathbf{A}'+\mathbf{B}'$ and both $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$ are symmetric, $\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}$ is symmetric.

Next, we show that $\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}$ is idempotent. First, note that $\mathbf{X}'\left(\mathbf{I}_n-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)=\mathbf{0}$. In partitioned form

$$\begin{bmatrix}\mathbf{X}_1'\\ \mathbf{X}_2'\end{bmatrix}\left(\mathbf{I}_n-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)=\mathbf{0}$$

which implies that

$$\mathbf{X}_1'\left(\mathbf{I}_n-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)=\mathbf{0}\ \text{ and }\ \mathbf{X}_2'\left(\mathbf{I}_n-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)=\mathbf{0}$$

so that

$$\mathbf{X}_1'=\mathbf{X}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \text{ and }\ \mathbf{X}_2'=\mathbf{X}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Now, keeping in mind that $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$ are idempotent , then

$$\begin{aligned}\mathbf{P}^2_{\mathbf{X}_1|\mathbf{X}_2}&=\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\right)\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\right)\\&=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'+\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\end{aligned}$$

Substituting $\mathbf{X}_2$ for $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2$ in the second term of the last line and $\mathbf{X}_2'$ for $\mathbf{X}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ in the third term,

$$\begin{aligned}\mathbf{P}^2_{\mathbf{X}_1|\mathbf{X}_2}&=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'+\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\\&=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'+\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\\&=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\\&=\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\end{aligned}$$

Thus $\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}$ is idempotent as claimed.

Finally,

$$\begin{aligned} rank\left(\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\right) &= trace\left(\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\right) \\ &= trace\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\right) \\ &= trace\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) - trace\left(\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\right) \\ &= q-(q-s) \\ &= s \end{aligned}$$

■

To confirm our choice of test statistic, we need to determine the expected values of the mean squares in the proposed test statistic. We already know the expected value of the denominator. The expected value of the numerator is given , without proof, by the next theorem.

Theorem II.20: Let $R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2) = \mathbf{Y}'\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{Y}$ be the reduction in sum of squares for adding the parameters $\boldsymbol{\theta}_1$ to a model that has parameters $\boldsymbol{\theta}_2$ where $\boldsymbol{\theta}_1$ is an $s\times 1$ column vector, $\boldsymbol{\theta}_2$ is an $(q-s)\times 1$ column vector, $\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}$ is a symmetric idempotent of rank s, $\mathbf{X}$ is an $n\times q$ matrix and $\mathbf{X}_2$ is an $n\times(q-s)$ matrix. Let $\mathbf{R}_{\mathbf{X}_2} = \mathbf{I}_n - \mathbf{P}_{\mathbf{X}_2} = \mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$. Then,

$$E\left[\mathbf{Y}'\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{Y}/s\right] = \sigma^2 + (1/s)\boldsymbol{\theta}_1'\mathbf{X}_1'\mathbf{R}_{\mathbf{X}_2}\mathbf{X}_1\boldsymbol{\theta}_1$$

with $\boldsymbol{\theta}_1'\mathbf{X}_1'\mathbf{R}_{\mathbf{X}_2}\mathbf{X}_1\boldsymbol{\theta}_1 > 0$ provided $\boldsymbol{\theta}_1 \neq \mathbf{0}$.

Proof: not given ■

So, under the null hypothesis H$_0$: $\boldsymbol{\theta}_1 = \mathbf{0}$, the expected value of the numerator mean square is σ^2 which is equal to the expected value of the Residual mean square. The test statistic will have a value of about one. If the hypothesis is not true, the expected value of the numerator mean square is greater than the Residual mean square and the ratio is expected to be larger than one. We can now determine the distribution of our proposed test statistic.

Theorem II.21: If H$_0$: $\boldsymbol{\theta}_1 = \mathbf{0}$ is true, then the test statistic

$$F_{s,(n-q)} = \frac{R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)/s}{D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)/(n-q)}$$

follows a Snedecor's F distribution with s and $(n-q)$ degrees of freedom.

Proof: To prove this result, we first need to use theorem II.16 to show that the two quadratic forms $\left(1/\sigma^2\right)R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)=\left(1/\sigma^2\right)\mathbf{Y}'\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{Y}$ and $\left(1/\sigma^2\right)D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=\left(1/\sigma^2\right)\mathbf{Y}'\mathbf{R}_\mathbf{X}\mathbf{Y}$ are independently distributed as chi-squares with s and $(n-q)$. degrees of freedom. To do this we have only to demonstrate that $E\left[\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{Y}\right]=E[\mathbf{R}_\mathbf{X}\mathbf{Y}]=\mathbf{0}$ because two of the three conditions listed in theorem II.16 are met — we already know that the two matrices are symmetric idempotents and it can be proved that $\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{R}_\mathbf{X}=\mathbf{0}$. We also know from theorems II.19 and II.10 that the ranks of $\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}$ and $\mathbf{R}_\mathbf{X}$ are s and $(n-q)$, respectively.

Firstly, note that given H$_0$: $\boldsymbol{\theta}_1=\mathbf{0}$,

$$E[\mathbf{Y}]=\mathbf{X}\boldsymbol{\theta}=[\mathbf{X}_1 \quad \mathbf{X}_2]\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\theta}_2\end{bmatrix}=\mathbf{X}_2\boldsymbol{\theta}_2.$$

Consequently,

$$\begin{aligned}E\left[\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{Y}\right]&=\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}E[\mathbf{Y}]\\ &=\left\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'-\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\right\}\mathbf{X}_2\boldsymbol{\theta}_2\\ &=\left\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2-\mathbf{X}_2\right\}\boldsymbol{\theta}_2\end{aligned}$$

In proving theorem II.19 it was shown that $\mathbf{X}_2'=\mathbf{X}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and so $\mathbf{X}_2=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2$. Hence,

$$\begin{aligned}E\left[\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{Y}\right]&=\left\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2-\mathbf{X}_2\right\}\boldsymbol{\theta}_2\\ &=\{\mathbf{X}_2-\mathbf{X}_2\}\boldsymbol{\theta}_2\\ &=\mathbf{0}\end{aligned}$$

As noted in proving theorem II.19, the expression $\mathbf{X}_2'=\mathbf{X}_2'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is equivalent to $\mathbf{X}_2'\left(\mathbf{I}_n-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)=\mathbf{0}$ so that $\mathbf{X}_2'\mathbf{R}_\mathbf{X}=\mathbf{R}_\mathbf{X}\mathbf{X}_2=\mathbf{0}$ and $E[\mathbf{R}_\mathbf{X}\mathbf{Y}]=\mathbf{R}_\mathbf{X}E[\mathbf{Y}]=\mathbf{R}_\mathbf{X}\mathbf{X}_2\boldsymbol{\theta}_2=\mathbf{0}$.

Now that we have proved the two quadratic forms $\left(1/\sigma^2\right)\mathbf{Y}'\mathbf{P}_{\mathbf{X}_1|\mathbf{X}_2}\mathbf{Y}$ and $\left(1/\sigma^2\right)\mathbf{Y}'\mathbf{R}_\mathbf{X}\mathbf{Y}$ are distributed independently as chi-squares with s and $(n-q)$ degrees of freedom, theorem II.17 is straightforwardly applied to conclude that

$$F_{s,(n-q)}=\frac{R(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2)/s}{D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)/(n-q)}$$

follows a Snedecor's F distribution with s and $(n-q)$ degrees of freedom as claimed.

■

Again our analysis can be conveniently summarized in an analysis of variance table.

Source	**DF**	**SSq**	**MSq**	**F**
Regression $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$	q	$R(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$		
Regn $\boldsymbol{\theta}_2$	q-s	$R(\boldsymbol{\theta}_2)$	$R(\boldsymbol{\theta}_2)/(q-s)$	
Regn $\boldsymbol{\theta}$ given $\boldsymbol{\theta}_2$	s	$R(\boldsymbol{\theta}_1 \vert \boldsymbol{\theta}_2)$	$R(\boldsymbol{\theta}_1 \vert \boldsymbol{\theta}_2)/s$	$\dfrac{R(\boldsymbol{\theta}_1 \vert \boldsymbol{\theta}_2)/s}{D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)/(n-q)}$
Residual	n-q	$D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$	$D(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)/(n-q)$	
Total	n	$\mathbf{Y}'\mathbf{Y}$		

e) Corrected sums of squares

One very important application of hypothesis testing of subvectors is where one tests for all coefficients other than the intercept are zero. As discussed previously, it is unusual for the intercept to be zero. Consequently, rather than testing the hypothesis H$_0$: $\boldsymbol{\theta} = \mathbf{0}$, we are more likely to want to test the hypothesis H$_0$: $\boldsymbol{\theta}_1 = \mathbf{0}$ where $\boldsymbol{\theta}_1$contains all parameters except the intercept; the intercept is the only parameter in $\boldsymbol{\theta}_2$. The main simplification that occurs is that $R(\boldsymbol{\theta}_2)$ is given by the following simple expression:

$$R(\boldsymbol{\theta}_2) = \left(\sum_{i=1}^{n} Y_i\right)^2 \bigg/ n$$

So a test of whether the p explanatory variables are required in the model is achieved using an analysis of variance such as just outlined. However, as there is no interest in the intercept why not remove it from the table altogether. If we do then the remaining sums of squares must sum to $\mathbf{Y}'\mathbf{Y} - R(\boldsymbol{\theta}_2)$ as in the following analysis of variance table:

Source	DF	SSq	MSq	F
Regression	p	$R(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)$ $= R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) - R(\boldsymbol{\theta}_2)$	$\dfrac{R(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)}{p}$	$\dfrac{R(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)/p}{D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/(n-p-1)}$
Residual	n-p-1	$D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ $= \mathbf{Y}'\mathbf{Y} - R(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$	$\dfrac{D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{n-p-1}$	
Total (corrected)	n-1	$\mathbf{Y}'\mathbf{Y} - R(\boldsymbol{\theta}_2)$		

Note that the total and regression sums of squares are both 'corrected' for $R(\boldsymbol{\theta}_2)$ and this quantity is called the **correction factor**. Also, the Residual sum of square remains unchanged and that it can be computed as the corrected Total minus the Regression sums of squares. It is this table that many computer packages produce — they neglect to label the Total as the corrected Total.

Example II.1 House price (continued)

The analysis of variance for the example is:

Source	DF	SSq	MSq	F
Regression	3	13143.904	4381.3013	93.12
Residual	2	94.096	47.0480	
Total	5	13238.000		

and

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

The correction factor is:

$$\left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n = (50+40+52+47+65)/5 = 254^2/5 \text{=} 12903.2$$

So the full analysis, incorporating a test for Age and Area given the intercept is in the model, is

Source	DF	SSq	MSq	F	p
Regression	3	13143.904	4381.3013		
Intercept	1	12903.200	12903.2000		
Explanatory	2	240.704	120.3520	2.5581	0.2810
Residual	2	94.096	47.0480		
Total	5	13238.000			

The analysis giving corrected sums of squares is:

Source	DF	SSq	MSq	F	p
Regression	2	240.704	120.352	2.5581	0.2810
Residual	2	94.096	47.048		
Total	4	334.800			

Clearly, the last table removes the clutter associated with the intercept, which we are not interested in.

II.F Likelihood ratio testing in regression

In section II.E the ANOVA method for formulating hypothesis tests was discussed. In this method the (corrected) total sum of squares is partitioned into a set of sums of squares that measure the variation from different sources, the source for a particular sum of squares being identified from the expected value of the corresponding mean square. The hypothesis tests are then performed by taking the ratios of appropriate pairs of mean squares, these ratios following Snedecor's F distribution.

An alternative procedure for formulating hypothesis tests is to form the ratio of appropriate likelihoods.

Definition II.12: Let $\mathbf{Y} = \left(Y_1, Y_2, \ldots, Y_n\right)'$ be a random vector that has joint probability distribution function $f\left(\mathbf{y}; \boldsymbol{\xi}\right)$ where $\mathbf{y}$ is a vector of observations from a random sample and $\boldsymbol{\xi} \in \boldsymbol{\Omega}$. Let $L\left(\boldsymbol{\xi}; \mathbf{y}\right) = f\left(\mathbf{y}; \boldsymbol{\xi}\right)$ be the likelihood function for $\mathbf{Y}$. The **generalized likelihood ratio statistic** for testing the null hypothesis H$_0$: $\boldsymbol{\xi} \in \boldsymbol{\Omega}_0$ against the alternative hypothesis H$_a$: $\boldsymbol{\xi} \in \boldsymbol{\Omega} - \boldsymbol{\Omega}_0$ is:

$$r\left(\mathbf{y}\right) = \frac{\max\limits_{\boldsymbol{\xi} \in \boldsymbol{\Omega}_0} L\left(\boldsymbol{\xi}; \mathbf{y}\right)}{\max\limits_{\boldsymbol{\xi} \in \boldsymbol{\Omega}} L\left(\boldsymbol{\xi}; \mathbf{y}\right)} = \frac{L\left(\tilde{\boldsymbol{\xi}}_0; \mathbf{y}\right)}{L\left(\tilde{\boldsymbol{\xi}}; \mathbf{y}\right)} = \frac{f\left(\mathbf{y}; \tilde{\boldsymbol{\xi}}_0\right)}{f\left(\mathbf{y}; \tilde{\boldsymbol{\xi}}\right)}$$

where $\tilde{\boldsymbol{\xi}}$ is the maximum likelihood estimate of $\boldsymbol{\xi}$ and $\tilde{\boldsymbol{\xi}}_0$ is the maximum likelihood estimates under the restriction that H$_0$ is true. ■

It can be shown that $0 \leq r\left(\mathbf{y}\right) \leq 1$. However, any monotonic function of $r\left(\mathbf{y}\right)$ is called 'a' likelihood ratio statistic.

Theorem II.22: Let $\mathbf{Y}$ be a normally distributed random vector representing a random sample with $E\left[\mathbf{Y}\right] = \mathbf{X}\boldsymbol{\theta}$ and $\operatorname{var}\left[\mathbf{Y}\right] = \sigma^2\mathbf{I}_n$ where $\mathbf{X}$ is an $n \times q$ matrix of full rank, $\boldsymbol{\theta}$ is a $q \times 1$ vector of unknown parameters and $n \geq q$. The generalized likelihood ratio statistic for testing H$_0$: $\boldsymbol{\theta} = \mathbf{0}$ versus H$_a$: $\boldsymbol{\theta} \neq \mathbf{0}$ is given by

$$r(\mathbf{y}) = \left\{ \frac{D(\tilde{\boldsymbol{\theta}})}{\mathbf{y}'\mathbf{y}} \right\}^{n/2}$$

where $D(\tilde{\boldsymbol{\theta}}) = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\theta}})$ is the Residual sum of squares computed using $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$.

Proof: From theorems II.7 and II.8, the maximum likelihood estimates of $\boldsymbol{\theta}$ and σ^2 are given by $\tilde{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\tilde{\sigma}_n^2 = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\theta}})/n = D(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)/n$. In outlining the proof of theorem it was shown that the likelihood function for $\mathbf{Y}$ is:

$$L(\boldsymbol{\xi}; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \right\}$$

Consequently, the likelihood function evaluated at the estimate for $\boldsymbol{\xi} \in \boldsymbol{\Omega}$ is given by

$$\begin{aligned} L(\tilde{\boldsymbol{\xi}}; \mathbf{y}) &= (2\pi\tilde{\sigma}^2)^{-n/2} \exp\left\{ \frac{-1}{2\tilde{\sigma}^2} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\theta}}) \right\} \\ &= \left\{ \frac{n}{2\pi D(\tilde{\boldsymbol{\theta}})} \right\}^{n/2} \exp\left\{ \frac{-n}{2D(\tilde{\boldsymbol{\theta}})} D(\tilde{\boldsymbol{\theta}}) \right\} \\ &= \left\{ \frac{n}{2\pi D(\tilde{\boldsymbol{\theta}})} \right\}^{n/2} e^{-n/2} \end{aligned}$$

Obtain the likelihood function for $\boldsymbol{\xi} \in \boldsymbol{\Omega}_0$ by setting $\boldsymbol{\theta} = \mathbf{0}$ in the original likelihood function:

$$L(\boldsymbol{\xi}_0; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-\mathbf{y}'\mathbf{y}}{2\sigma^2} \right\}$$

In this case, the maximum likelihood estimate of $\tilde{\sigma}_n^2$ is $\tilde{\sigma}_n^2 = \mathbf{y}'\mathbf{y}/n$ and so the likelihood function evaluated at the estimate for $\boldsymbol{\xi} \in \boldsymbol{\Omega}_0$ is given by

$$
\begin{aligned}
L\left(\tilde{\boldsymbol{\xi}}_0;\mathbf{y}\right) &= \left(2\pi\tilde{\sigma}^2\right)^{-n/2}\exp\left\{\frac{-\mathbf{y}'\mathbf{y}}{2\tilde{\sigma}^2}\right\} \\
&= \left\{\frac{n}{2\pi\mathbf{y}'\mathbf{y}}\right\}^{n/2}\exp\left\{\frac{-n\mathbf{y}'\mathbf{y}}{2\mathbf{y}'\mathbf{y}}\right\} \\
&= \left\{\frac{n}{2\pi\mathbf{y}'\mathbf{y}}\right\}^{n/2} e^{-n/2}
\end{aligned}
$$

Hence the generalized ratio statistic is given by

$$
\begin{aligned}
r(\mathbf{y}) &= \frac{L\left(\tilde{\boldsymbol{\xi}}_0;\mathbf{y}\right)}{L\left(\tilde{\boldsymbol{\xi}};\mathbf{y}\right)} \\
&= \frac{\left\{\dfrac{n}{2\pi\mathbf{y}'\mathbf{y}}\right\}^{n/2} e^{-n/2}}{\left\{\dfrac{n}{2\pi D\left(\tilde{\boldsymbol{\theta}}\right)}\right\}^{n/2} e^{-n/2}} \\
&= \left\{\frac{D\left(\tilde{\boldsymbol{\theta}}\right)}{\mathbf{y}'\mathbf{y}}\right\}^{n/2}
\end{aligned}
$$

■

Next we show that the F statistic derived by the ANOVA method in section II.E is a monotonic function of the generalized likelihood ratio statistic just derived and so the F statistic is a likelihood ratio statistic.

Theorem II.23: Let **Y** be a normally distributed random vector representing a random sample with $E[\mathbf{Y}]=\mathbf{X\theta}$ and $\text{var}[\mathbf{Y}]=\sigma^2\mathbf{I}_n$ where **X** is an $n\times q$ matrix of full rank, $\boldsymbol{\theta}$ is a $q\times 1$ vector of unknown parameters and $n \geq q$. The F statistic for testing the hypothesis H$_0$: $\boldsymbol{\theta} = \mathbf{0}$ versus H$_a$: $\boldsymbol{\theta} \neq \mathbf{0}$ is a monotonic function of the likelihood ratio statistic for this test.

Proof: The F statistic is

$$
F_{q,\,n-q} = \frac{R(\boldsymbol{\theta})/q}{D(\boldsymbol{\theta})/(n-q)}
$$

where $R(\boldsymbol{\theta})$ is the Regression sums of squares.

Recall that $\mathbf{y}'\mathbf{y} = R\left(\tilde{\boldsymbol{\theta}}\right)+D\left(\tilde{\boldsymbol{\theta}}\right)$ so that the likelihood ratio statistic can be re-expressed as

$$
\begin{aligned}
r(\mathbf{y}) &= \left\{\frac{D(\tilde{\boldsymbol{\theta}})}{\mathbf{y}'\mathbf{y}}\right\}^{n/2} \\
&= \left\{\frac{D(\tilde{\boldsymbol{\theta}})}{R(\tilde{\boldsymbol{\theta}})+D(\tilde{\boldsymbol{\theta}})}\right\}^{n/2} \\
&= \left\{\frac{1}{\dfrac{R(\tilde{\boldsymbol{\theta}})}{D(\tilde{\boldsymbol{\theta}})}+1}\right\}^{n/2} \\
&= \left\{1+\frac{q\left\{R(\tilde{\boldsymbol{\theta}})/q\right\}}{(n-q)\left\{D(\tilde{\boldsymbol{\theta}})/(n-q)\right\}}\right\}^{-n/2} \\
&= \left\{1+\frac{q}{(n-q)}F_{q,n-q}\right\}^{-n/2}
\end{aligned}
$$

Solving this equation for $F_{q,n\text{-}q}$ yields

$$F_{q,n-q} = \frac{(n-q)}{q}\left\{r(\mathbf{y})^{-2/n}-1\right\}$$

and the derivative of $F_{q,n\text{-}q}$ with respect to $r(\mathbf{y})$ is

$$-\frac{2(n-q)}{nq}r(\mathbf{y})^{-(2/n)-1}$$

Now $0 \le r(\mathbf{y}) \le 1$ and the derivative is always negative for $0 < r(\mathbf{y}) < 1$ so it is a strictly decreasing function. ■

The import of this last theorem is that a test based on the likelihood ratio statistic will produce the same results as on based one the $F_{q,n\text{-}q}$ test statistic.

II.G Summary

In this chapter we have:

- introduced the linear model used in regression including an alternative expression for it in terms of an expectation and variance model;
- derived the least squares estimators of the expectation model parameters and of linear functions of them, showing that the estimators are BLUE and mentioned that they are also UMVUE;
- obtained an estimator, not the maximum likelihood estimator, for the variance and demonstrated that it is unbiased;
- demonstrated that the maximum likelihood estimators of the expectation model parameters are identical to the least squares estimators;
- outlined an hypothesis test for testing that expectation model parameters are zero; this test is an analysis of variance with Source, DF, Sq, MSq and F values for each line in the table; it involves:
    1. formulating a partition of the total sums of squares into regression and residual sums of squares with accompanying degrees of freedom; in particular showed that:
        - ⇒ the sums of squares can be expressed as quadratic forms of the observation vector **y** with the matrix of the quadratic form being one of the idempotents **P** or **R**;
        - ⇒ the degrees of freedom for a term are equal to the trace of the idempotent;
    2. computing the mean squares from the sums of squares and degrees of freedom;
    3. deriving the expected mean squares — the mean value of the mean squares under sampling from a population in which the variables *X* and *Y* are related as specified by a linear regression model; they allow us to determine an F statistic for testing the null hypothesis;
    4. computing the F statistic and showing that its sampling distribution under the null hypothesis is a Snedecor's F distribution.
- Examined an hypothesis test for testing that a subset of the expectation model parameters is zero; the reduction in the Residual sum of squares resulting from the inclusion of the subset in the model was specified to be the difference in the residual sums of squares or deviances obtained from fitting models that include and do not include the subset;
- applied the hypothesis test for a subset of parameters to a test for all parameters except the intercept being zero;
- proved that the F test statistics is equivalent to a likelihood ratio test.