MAS programmes - Statistical Analysis (Autumn Term)

2 Multiple Linear Regression

2.1 The multiple linear regression model

In multiple linear regression, a response variable Y is expressed as a linear function of k regressor (or *predictor* or *explanatory*) variables, $X_1, X_2, \ldots, X_k$, with corresponding model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon.$$

Suppose now that n observations have been made of the variables (where $n \geq k+1$) with x_{ij} the ith observed value of the jth regressor variable, corresponding to the ith observed value y_i of the response variable. Thus the data could be tabulated in the following form of a data matrix.

Observation	Variable				
Number	y	x_1	x_2	$\ldots$	x_k
1	y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

The *multiple linear regression model* is

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \qquad i = 1, \ldots, n \tag{2.1}$$

where the x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$, are regarded as fixed, $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters and the errors ϵ_i, $i = 1, \ldots, n$, are assumed to be i.i.d.$(0, \sigma^2)$, with σ^2 unknown.

The model (2.1) may be written in matrix notation as

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations,

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T,$$

$\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, where $p = k + 1$,

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T,$$

$\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors,

$$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T,$$

and $\mathbf{X}$ is an $n \times p$ matrix, the *design matrix*,

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

Equation (2.2) expresses the regression model in the form of what is known as the *general linear model.*

Note that in this form, we have

$$\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \quad \text{and} \quad \mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

It follows that $\mathrm{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\mathbf{y}, \mathbf{y}) = \sigma^2\mathbf{I}$. Note that, as in the case of simple linear regression, we have so far made no assumptions about the distribution of the y_i.

2.2 Estimation of Parameters

According to the *method of least squares*, we choose as our estimates of the vector of parameters $\boldsymbol{\beta}$ the vector $\mathbf{b} = (b_0, b_1, \dots, b_k)^T$ whose elements jointly minimize the functional

$$\mathcal{L} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.3}$$

i.e.,

$$\mathcal{L} = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} x_{ij}\beta_j \right)^2$$

where $x_{i0}=1,\ i=1,\dots,n$. This expression is minimized by setting the partial derivatives with respect to each of the $\beta_r,\ r=0,\dots,k$, equal to zero. This yields the *normal equations*, a set of $p = k+1$ simultaneous linear equations for the p unknowns, $b_0, b_1, \dots, b_k$,

$$\sum_{i=1}^{n}\sum_{j=0}^{k} x_{ir}x_{ij}b_j = \sum_{i=1}^{n} x_{ir}y_i \qquad r=0,\dots,k$$

which may be written in matrix form as

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}. \tag{2.4}$$

To see this directly in matrix algebra terms, write (2.3) as

$$\begin{aligned} \mathcal{L} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Then,

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right)$$
$$= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

(using results (F.5) and (F.7) of the *Notes for MSc Students*), which, when evaluated at $\boldsymbol{\beta} = \mathbf{b}$, results in the normal equations given above.

Note that $\mathbf{X}^T\mathbf{X}$ is a symmetric $p \times p$ matrix. Also, the minimum value obtained for the functional $\mathcal{L}$, evaluated at $\mathbf{b}$, is the *error (or residual) sum of squares* SS_R.

2.3 Rank and invertibility

The *rank*, $\text{rank}(\mathbf{A})$, of a matrix $\mathbf{A}$ is the number of linearly independent columns of $\mathbf{A}$.

Recall that our design matrix $\mathbf{X}$ is an $n \times p$ matrix with $n \geq p$. It follows that $\text{rank}(\mathbf{X}) \leq p$. If $\text{rank}(\mathbf{X}) = p$ then $\mathbf{X}$ is said to be of *full rank*. It may be shown that

$$\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}). \tag{2.5}$$

A square matrix is said to be *non-singular* if it has an inverse. A $p \times p$ square matrix is non-singular if and only if it is of full rank p.

If $\mathbf{X}^T\mathbf{X}$ is non-singular, which by the result (2.5) occurs if and only if $\mathbf{X}$ is of full rank p, then the normal equations (2.4) have a unique solution,

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.6}$$

It will generally be the case for sensible regression models that the design matrix $\mathbf{X}$ is of full rank, but this is not necessarily always the case. To take an extreme example, if one of the regressor variables is a scaled version of another then the two corresponding columns of the matrix $\mathbf{X}$ are scalar multiples of each other and hence $\text{rank}(\mathbf{X}) < p$. The normal equations do not then have a unique solution – the estimates of the parameters are not well-determined.

For a given set of data, assuming that $\mathbf{X}$ is of full rank p, the formal mathematical solution (2.6) of the normal equations (2.4) is translated in a statistical package such as R into a numerical procedure for solving the normal equations.

2.4 The hat matrix

Assume that $\mathbf{X}$ is of full rank. The vector $\hat{\mathbf{y}}$ of fitted values is given by

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{y}, \tag{2.7}$$

where, using Equation (2.6), the *hat matrix* $\mathbf{H}$ is defined by

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T. \tag{2.8}$$

Note that $\mathbf{H}$ is a symmetric $n \times n$ matrix. The vector $\mathbf{e}$ of residuals is given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}, \tag{2.9}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

Before going any further it is helpful to note some results about the matrices $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$. A matrix $\mathbf{P}$ is said to be *idempotent* if $\mathbf{P^2} = \mathbf{P}$.

From Equation (2.8),

$$\begin{aligned} \mathbf{H^2} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{H}. \end{aligned} \tag{2.10}$$

Thus $\mathbf{H}$ is idempotent. Furthermore, using Equation (2.10),

$$(\mathbf{I} - \mathbf{H})^{\mathbf{2}} = \mathbf{I^2} - \mathbf{2H} + \mathbf{H^2} = \mathbf{I} - \mathbf{H}. \tag{2.11}$$

Thus $\mathbf{I} - \mathbf{H}$ is also an $n \times n$ symmetric idempotent matrix. Again using Equation (2.10),

$$\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{O}_{n \times n}, \tag{2.12}$$

where $\mathbf{O}_{n \times n}$ represents a matrix of zeros, in this case an $n \times n$ matrix.

Post-multiplying Equation (2.8) by $\mathbf{X}$, we find that

$$\mathbf{HX} = \mathbf{X}. \tag{2.13}$$

Hence

$$(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{O}_{n \times p}, \tag{2.14}$$

where $\mathbf{O}_{n \times p}$ again represents a matrix of zeros, but now an $n \times p$ matrix.

Recall that the trace of a matrix is the sum of its diagonal elements.

$$\begin{aligned} \mathrm{tr}(\mathbf{H}) &= \mathrm{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\ &= \mathrm{tr}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}) \\ &= \mathrm{tr}(\mathbf{I}_p) = p, \end{aligned}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix. It follows that

$$\mathrm{tr}(\mathbf{I} - \mathbf{H}) = \mathrm{tr}(\mathbf{I}) - \mathrm{tr}(\mathbf{H}) = n - p.$$

It turns out that the rank of a symmetric idempotent matrix is equal to its trace. Hence

$$\mathrm{rank}(\mathbf{H}) = p$$

and

$$\mathrm{rank}(\mathbf{I} - \mathbf{H}) = n - p.$$

2.5 Properties of the least squares estimator

In the linear model as specified in Equation (2.2), $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. It follows that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\mathbf{y}, \mathbf{y}) = \sigma^2\mathbf{I}$. We now consider the properties of the least squares estimator $\mathbf{b}$ as specified in Equation (2.6). (For the present, we do not need to use the normality assumption for the error distribution.)

$$\begin{aligned}
\mathrm{E}(\mathbf{b}) &= \mathrm{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}(\mathbf{y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}. \qquad (2.15)
\end{aligned}$$

Thus $\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\beta}$. The covariance matrix of $\mathbf{b}$ is found as follows.

$$\begin{aligned}
\mathrm{Cov}(\mathbf{b}, \mathbf{b}) &= \mathrm{Cov}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Cov}(\mathbf{y}, \mathbf{y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}. \qquad (2.16)
\end{aligned}$$

Theorem 2.1 (*The Gauss-Markov Theorem*)
For any $p \times 1$ vector $\mathbf{a}$, $\mathbf{a}^T\mathbf{b}$ is the unique minimum variance linear unbiased estimator of $\mathbf{a}^T\boldsymbol{\beta}$.
Proof Let $\mathbf{c}^T\mathbf{y}$ be any linear unbiased estimator of $\mathbf{a}^T\boldsymbol{\beta}$. It follows that, for all $\boldsymbol{\beta}$,

$$E(\mathbf{c}^T\mathbf{y}) = \mathbf{c}^T\boldsymbol{X}\boldsymbol{\beta} = \mathbf{a}^T\boldsymbol{\beta}.$$

Hence it must be the case that $\mathbf{c}^T\mathbf{X} = \mathbf{a}^T$. The variance of the estimator $\mathbf{c}^T\mathbf{y}$ is given by

$$\mathrm{var}(\mathbf{c}^T\mathbf{y}) = \mathbf{c}^T\sigma^2\mathbf{I}\mathbf{c} = \sigma^2\mathbf{c}^T\mathbf{c}.$$

Now consider the properties of the estimator $\mathbf{a}^T\mathbf{b}$. Firstly, it is unbiased, since

$$E(\mathbf{a}^T\mathbf{b}) = \mathbf{a}^T E(\mathbf{b}) = \mathbf{a}^T\boldsymbol{\beta},$$

using Equation (2.15). Using Equation (2.16), the variance of $\mathbf{a}^T\mathbf{b}$ is given by

$$\begin{aligned}
\mathrm{var}(\mathbf{a}^T\mathbf{b}) &= \mathbf{a}^T\mathrm{cov}(\mathbf{b}, \mathbf{b})\mathbf{a} \\
&= \mathbf{a}^T\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} \\
&= \sigma^2\mathbf{c}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{c} \\
&= \sigma^2\mathbf{c}^T\mathbf{H}\mathbf{c},
\end{aligned}$$

where $\mathbf{H}$ is the hat matrix as defined in Equation (2.8). Using the fact that $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent, it follows that

$$\begin{aligned}
\mathrm{var}(\mathbf{c}^T\mathbf{y}) - \mathrm{var}(\mathbf{a}^T\mathbf{b}) &= \sigma^2\mathbf{c}^T(\mathbf{I} - \mathbf{H})\mathbf{c} \\
&= \sigma^2((\mathbf{I} - \mathbf{H})\mathbf{c})^T((\mathbf{I} - \mathbf{H})\mathbf{c}) \\
&= \sigma^2\|(\mathbf{I} - \mathbf{H})\mathbf{c}\|^2 \\
&\geq 0,
\end{aligned}$$

with equality if and only if $\mathbf{c} = \mathbf{H}\mathbf{c}$. This is true if and only if, for all $\mathbf{y}$,

$$\begin{aligned}
\mathbf{c}^T\mathbf{y} &= \mathbf{c}^T\mathbf{H}\mathbf{y} \\
&= \mathbf{c}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{a}^T\mathbf{b}.
\end{aligned}$$

2.6 The estimation of the error variance

Lemma Let $\mathbf{y}$ be an $n \times 1$ vector of random variables and $\mathbf{A}$ an $n \times n$ symmetric matrix of constants. If $\mathrm{E}(\mathbf{y}) = \boldsymbol{\theta}$ and $\mathrm{Cov}(\mathbf{y}, \mathbf{y}) = \boldsymbol{\Sigma}$ then

$$E(\mathbf{y}^T\mathbf{A}\mathbf{y}) = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}.$$

Proof

$$\begin{aligned}
\mathrm{E}(\mathbf{y}^T\mathbf{A}\mathbf{y}) &= \mathrm{E}\{(\mathbf{y}-\boldsymbol{\theta})^T\mathbf{A}(\mathbf{y}-\boldsymbol{\theta}) + 2\boldsymbol{\theta}^T\mathbf{A}\mathbf{y} - \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}\} \\
&= \mathrm{E}\{(\mathbf{y}-\boldsymbol{\theta})^T\mathbf{A}(\mathbf{y}-\boldsymbol{\theta})\} + 2\boldsymbol{\theta}^T\mathbf{A}\mathrm{E}(\mathbf{y}) - \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta} \\
&= \sum_i\sum_j a_{ij}\mathrm{E}\{(y_i-\theta_i)(y_j-\theta_j)\} + \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta} \\
&= \sum_i\sum_j a_{ij}\sigma_{ij} + \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta} \\
&= \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}.
\end{aligned}$$

Recalling Equations (2.9) and (2.11), we apply the result of the lemma to the residual sum of squares,

$$\mathbf{e}^T\mathbf{e} = \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y},$$

with $\mathbf{A} = \mathbf{I} - \mathbf{H}$, $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. Thus

$$\mathrm{E}(\mathbf{e}^T\mathbf{e}) = \sigma^2\mathrm{tr}(\mathbf{I}-\mathbf{H}) + \boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta}.$$

From Section 2.4, $\mathrm{tr}(\mathbf{I}-\mathbf{H}) = n-p$. Using Equation (2.14), the second term on the right hand side of the above equation is zero. Hence

$$E(\mathbf{e}^T\mathbf{e}) = (n-p)\sigma^2.$$

Thus an unbiased estimator of the error variance σ^2 is given by the residual mean square (MS_R),

$$s^2 \equiv \frac{\mathbf{e}^T\mathbf{e}}{n-p} = \frac{SS_R}{n-p} = \frac{(\mathbf{y}-\mathbf{X}\mathbf{b})^T(\mathbf{y}-\mathbf{X}\mathbf{b})}{n-p}. \tag{2.17}$$

2.7 The normality assumption

To this point we have proceeded without any assumptions about the distribution of the response variable Y_i; we have assumed only that they are independent and have constant variance. The model is

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n$$

where $\mathbf{x}_i^T$ is the ith row of the design matrix $\mathbf{X}$ corresponding to the observations on the ith individual, and the errors $\epsilon_i \sim$ i.i.d.$(0, \sigma^2)$. More generally, the model can be written

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.18}$$

The method of least squares has provided us with estimates of the regression coefficients (parameter vector) $\boldsymbol{\beta}$, *viz*

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.19}$$

which are linear combinations of the dependent variables. We have discussed the properties of the *ordinary least squares* (OLS) estimator $\mathbf{b}$ and obtained expressions for its expected value and covariance matrix.

$$\mathrm{E}[\mathbf{b}] = \boldsymbol{\beta} \tag{2.20}$$

$$\mathrm{Cov}(\mathbf{b}, \mathbf{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \tag{2.21}$$

Suppose now that normality can be assumed for the Y_i, so that the multiple linear regression model can be written

$$Y_i \sim \mathrm{NID}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$$

or equivalently,

$$\mathbf{Y} \sim \mathrm{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \tag{2.22}$$

This is equivalent to the assumption that $\boldsymbol{\epsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\mathbf{I})$ in (2.18).

Using the normality assumption, the estimates of $\boldsymbol{\beta}$ may be obtained alternatively using the method of maximum likelihood. Given $\mathbf{Y}$ and $\mathbf{X}$, the likelihood may be written

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= |2\pi\sigma^2\mathbf{I}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\} \end{aligned}$$

so that the log-likelihood function is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

and hence

$$S(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}\{2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{y}\}.$$

Equating this expression to zero, leads us directly to the normal equations seen earlier. Note that $S(\boldsymbol{\beta})$ is often referred to as the *score* function.

Hence it is seen that the OLS estimator of $\boldsymbol{\beta}$, $\mathbf{b}$, is the same as the maximum likelihood estimator (MLE), under the normality assumption, i.e. $\hat{\boldsymbol{\beta}} = \mathbf{b}$.

Further, the MLE of σ^2 can be found from

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2}\right) + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

and, equating to zero gives

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n} = \frac{\mathbf{e}^T\mathbf{e}}{n} \tag{2.23}$$

so that the maximum likelihood estimator of σ^2 is biased. For this reason the unbiased estimator

$$s^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-p}$$

seen in the previous section is preferred. However, as $n \longrightarrow \infty$ the bias of the maximum likelihood estimator shrinks towards zero, so that the MLE is consistent.

Under the assumption of normality, it can be shown that

$$\mathbf{b} \sim \mathrm{MVN}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \tag{2.24}$$

and

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p} \tag{2.25}$$