MA 575 Linear Models:

Cedric E. Ginestet, Boston University

Maximum Likelihood Estimation

Week 6, Lecture 1

1 The Multivariate Normal Distribution

1.1 Probability Density Function

The multidimensional equivalent of the normal or **Gaussian** distribution is defined for a random vector, $\mathbf{x} \in \mathbb{R}^p$, of the form $\mathbf{x} := [X_1, \ldots, X_p]^T$,

$$\mathbf{x} \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1}$$

where $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_p]^T$ is the mean of $\mathbf{x}$, and $\boldsymbol{\Sigma}$ is a *positive definite* $p \times p$ covariance matrix. Using our adopted notation for the moments of random vectors, we have

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \qquad \text{and} \qquad \mathbb{V}\text{ar}[\mathbf{x}] = \boldsymbol{\Sigma}.$$

The corresponding **joint** probability density function (pdf) for $\mathbf{x} = [X_1, \ldots, X_p]^T$ is given by

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\},$$

with $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. The **marginal** and **conditional** variances of the X_i's are obtained as follows,

i. The **marginal** variance of X_i is $\mathbb{V}\text{ar}[X_i] = \boldsymbol{\Sigma}_{ii}$,

ii. The **conditional** variance of $X_i|\mathbf{X}_{-i}$ is $\mathbb{V}\text{ar}[X_i|\mathbf{X}_{-i}] = \mathbf{Q}_{ii}^{-1}$,

where $\mathbf{Q} := \boldsymbol{\Sigma}^{-1}$ is commonly referred to as the **precision** matrix. The properties of the bivariate normal distribution can easily be studied visually using a scatterplot.

Remark 1. The condition that $\boldsymbol{\Sigma} > 0$ means that for every non-zero vector $\mathbf{v} \in \mathbb{R}^p$, we have $\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} > 0$. Positive definiteness is a stronger condition than $|\boldsymbol{\Sigma}| > 0$, since every positive definite matrix has a positive determinant, but not all matrices with a positive determinant are positive definite.

1.2 Bivariate Normal Distribution

Let us return to our simple regression example and assume that both Y and X are random, and are **jointly** distributed as a bivariate normal distribution, such that

$$\begin{vmatrix} x_i \\ y_i \end{vmatrix} \sim \text{MVN}_2 \left(\begin{vmatrix} \mu_x \\ \mu_y \end{vmatrix}, \begin{vmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho_{xy} \\ \sigma_x \sigma_y \rho_{xy} & \sigma_y^2 \end{vmatrix} \right).$$

Now, from the standard properties of the Normal distribution, we can readily obtain the **marginal distributions** of both Y and X, as follows,

$$y_i \sim N(\mu_y, \sigma_y^2) \qquad \text{and} \qquad x_i \sim N(\mu_x, \sigma_x^2).$$

In addition, the **conditional distribution** of Y, given X, which is especially relevant for understanding simple linear regression is given by

$$y_i|x_i \sim N\bigg(\mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x), \sigma_y^2(1-\rho_{xy}^2)\bigg).$$

Clearly, one can immediately recognize the standard mean and variance functions of simple linear regression. To make this relationship transparent, we define

$$\beta_0 := \mu_y - \beta_1\mu_x, \qquad \beta_1 := \rho_{xy}\frac{\sigma_y}{\sigma_x}, \qquad \text{and} \qquad \sigma^2 := \sigma_y^2(1-\rho_{xy}^2),$$

in order to obtain

$$y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma_{y|x}^2).$$

This is quite remarkable. Our definition of simple linear regression exactly coincide with the specification of a joint normal distribution for the observations y_i's, and the predictor x_i's. In addition, recall that we have found that the *coefficient of determination* for simple linear regression is directly equivalent to the square of the estimated correlation between X and Y. Therefore, we can re-write the conditional variance as

$$\sigma_{y|x}^2 = \sigma_y^2(1-R^2),$$

where R^2 is here defined as the square of the **population** correlation between X and Y. That is, the greater is the statistical dependence between X and Y, the smaller is the conditional variance, $\sigma_{y|x}$. That is, if these two random variables are perfectly correlated, then X entirely predicts Y.

1.3 Multiple Regression

Here, we ignore the intercept, and simply consider the relationship between the p predictors and the Y. Moreover, we assume that have a multivariate Normal distribution of the following form,

$$\begin{vmatrix}\mathbf{x}_i \\ y_i\end{vmatrix} \sim \left(\begin{vmatrix}\boldsymbol{\mu}_x \\ \mu_y\end{vmatrix}, \begin{vmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \sigma_y^2\end{vmatrix}\right),$$

where $\boldsymbol{\Sigma}_{yx} := \boldsymbol{\Sigma}_{xy}^T$ and $\boldsymbol{\Sigma}_{xx}$ is a $p \times p$ matrix of covariances between the elements of $\mathbf{x}_i$. Moreover, $\boldsymbol{\Sigma}_{xy}$ is a $p \times 1$ vector of covariances between y_i and $\mathbf{x}_i$, such that this covariance matrix has the following form,

$$\boldsymbol{\Sigma} = \begin{vmatrix} \mathbb{V}\text{ar}[X_1] & \dots & \mathbb{C}\text{ov}[X_1, X_p] & \mathbb{C}\text{ov}[X_1, Y] \\ \vdots & \ddots & \ddots & \vdots \\ \mathbb{C}\text{ov}[X_p, X_1] & \dots & \mathbb{V}\text{ar}[X_p] & \mathbb{C}\text{ov}[X_p, Y] \\ \mathbb{C}\text{ov}[Y, X_1] & \dots & \mathbb{C}\text{ov}[Y, X_p] & \sigma_y^2 \end{vmatrix}.$$

Moreover, the multivariate normal distribution also provides the **conditional** distribution of y_i given $\mathbf{x}_i$,

$$y_i|\mathbf{x}_{-i} \sim N\Big(\mu_x + \boldsymbol{\beta}^{*T}(\mathbf{x}_i - \boldsymbol{\mu}_x),\ \sigma_y^2 - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\Big). \tag{2}$$

Here, the conditional variance of X_i given $\mathbf{X}_{-1}$ can be obtained by considering the inverse of $\boldsymbol{\Sigma}$, denoted $\mathbf{Q} := \boldsymbol{\Sigma}^{-1}$. Moreover, the set of coefficients

$$\boldsymbol{\beta}^* := \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy},$$

is related to $\widehat{\boldsymbol{\beta}}^* := (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}\mathcal{Y}$, as described in your textbook, on p.56.

1.4 Schur Complement

Recall that the inverse of a 2×2 matrix $\mathbf{M}$ is given by

$$\mathbf{M}^{-1} = \left(\begin{vmatrix} a & b \\ c & d \end{vmatrix} \right)^{-1} = \frac{1}{ad - bc} \begin{vmatrix} d & -b \\ -c & a \end{vmatrix},$$

where the first element of this inverse can be written as

$$(\mathbf{M}^{-1})_{11} := \left(a - \frac{bc}{d} \right)^{-1}.$$

Similarly, for a **block matrix** as above, such as

$$\boldsymbol{\Sigma} := \begin{vmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{vmatrix},$$

the first element of the inverse matrix $\mathbf{Q} := \boldsymbol{\Sigma}^{-1}$ is given by

$$\mathbf{Q}_{11} = \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \right)^{-1}.$$

The quantity, $\mathbf{Q}_{11}^{-1}$ as used in equation (2), is important in its own right, and is generally referred to as the **Schur complement** of $\boldsymbol{\Sigma}_{22}$ in $\boldsymbol{\Sigma}$.

1.5 Relationships between Predictors

Recall that the distribution of the OLS estimators $\widehat{\boldsymbol{\beta}}$ also follows a multivariate normal distribution, such that

$$\widehat{\boldsymbol{\beta}} \sim \text{MVN}_{p^*}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}).$$

We have also seen that the marginal distribution of each $\widehat{\beta}_j$ is straightforwardly given by

$$\widehat{\beta}_j \sim N(\beta_j, \sigma^2 [(\mathbf{X}^T \mathbf{X})^{-1}]_{jj}),$$

for every $j = 1, \ldots, p^*$. For presentational convenience, let us restrict our attention to the case $p^* = 2$. If we consider one of these entries in this vector, say $j = 1$, we can re-write the **uncorrected sum of squares**, $\mathbf{S} := \mathbf{X}^T \mathbf{X}$,

$$\mathbf{S} := \mathbf{X}^T \mathbf{X} = \begin{vmatrix} s_{11}^2 & s_{12}^2 \\ s_{21}^2 & s_{22}^2 \end{vmatrix},$$

where $s_{11}^2 := (\mathbf{X}^T \mathbf{X})_{11}$. Using the Schur complement, the variance of this marginal distribution can be re-written as follows

$$\mathbb{V}\text{ar}[\widehat{\beta}_1 | \mathbf{X}] = \frac{\sigma^2}{s_{11}^2 - s_{12}^2 s_{22}^{-2} s_{21}^2},$$

The denominator should here be interpreted as the proportion of the variance of $\widehat{\beta}_1$, which is not explained by the remaining predictors.

2 Maximum Likelihood Estimation

We have just seen that the MLE and OLS estimators for $\boldsymbol{\beta}$ perfectly coincide in the case of multiple regression under normal errors. This is the case because we are, in fact, optimizing the same functional of the observed values.

2.1 The Likelihood Function

For some set of **independent** observations $(y_i, \mathbf{x}_i)$, with $i = 1, \ldots, n$, we assume the following **probabilistic model**,

$$y_i \overset{\text{ind}}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2), \qquad \forall\ i = 1, \ldots, n.$$

The likelihood function for this data set **parametrized** by $(\boldsymbol{\beta}, \sigma^2)$ is then defined as a product of densities,

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) := \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2).$$

In the case of multiple regression, the definition of the Normal distribution gives the following product,

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\}.$$

Intuitively, the **maximum likelihood estimator (MLE)** is defined as the *parameter value for which the data sample is the most likely.* For a linear model like multiple regression, the set of parameters, whose values need to be optimized are composed of the vector of coefficients $\boldsymbol{\beta}$ and the variance σ^2, such that the MLEs is a vector of the form

$$\widehat{\boldsymbol{\theta}}_{\text{MLE}} := (\widehat{\beta}_0, \ldots, \widehat{\beta}_p, \widehat{\sigma}^2).$$

One of the key properties of the normal distribution is that selecting the mean that maximize the likelihood of the data, is in some sense "**orthogonal**" of selecting the variance that maximizes the likelihood of the data. This particular notion of orthogonality will be made more precise in section 2.5.

2.2 Intercept and Slope Parameters

Since the logarithm is a **strictly monotonic** function, the argument maximizing $L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{x})$ is equivalent to the one maximizing its logarithm. Formally, we have

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{p^*}}{\operatorname{argmax}}\, L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p^*}}{\operatorname{argmax}}\, \log L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}),$$

and similarly for the MLE of the variance, σ^2. Observe that this particular equivalence would not hold, however, if the logarithmic transformation was solely *monotonic*, as this would not preserve the uniqueness of a local maximizer. Thus, the likelihood function for multiple regression can be simplified by noting that the *log of a product* is the *product of the logs.*

$$\begin{aligned} \log L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\} \right) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2. \end{aligned}$$

Finally, we can omit the first term in the log-likelihood since this does not depend on $\boldsymbol{\beta}$. This gives us an expression, which is strongly related to the **residual sum of squares**.

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

As before for the OLS estimators, we proceed by setting the first derivative of $L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X})$ to zero, and solving for $\boldsymbol{\beta}$, which yields

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial \boldsymbol{\beta}} \Big(\mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T (\mathbf{X}^T \mathbf{X}) \boldsymbol{\beta} \Big) \\ &= -\frac{1}{2\sigma^2} \Big(2(\mathbf{X}^T \mathbf{X}) \boldsymbol{\beta} - 2\mathbf{X}^T \mathbf{y} \Big), \end{aligned} \tag{3}$$

which yields $(\mathbf{X}^T \mathbf{X})\boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$. Therefore, this implies that assuming the normality of the error terms in multiple regression produces a complete equivalence between **maximizing the likelihood** and **minimizing the RSS**, such that

$$\widehat{\boldsymbol{\beta}}_{\text{MLE}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{p^*}}{\operatorname{argmax}} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}, \mathbf{X}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p^*}}{\operatorname{argmin}} \operatorname{RSS}(\boldsymbol{\beta}) =: \widehat{\boldsymbol{\beta}}_{\text{OLS}}.$$

2.3 Variance Parameter

We have seen that we can exploit the orthogonality of $\boldsymbol{\beta}$ and σ^2 in a Normal model, in order to maximize the likelihood by selecting these two sets of parameters independently of each other. Thus, once we have chosen $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$, it suffices to select

$$\widehat{\sigma}^2_{\text{MLE}} := \underset{\sigma^2 \in \mathbb{R}^+}{\operatorname{argmax}} \log L(\widehat{\boldsymbol{\beta}}_{\text{MLE}}, \sigma^2; \mathbf{y}, \mathbf{X}),$$

which gives

$$-\frac{\partial}{\partial \sigma^2} \left(\frac{n}{2} \log(2\pi) + \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{\text{MLE}})^2 \right) = 0.$$

This can be readily solved in order to obtain

$$\widehat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \operatorname{RSS}(\widehat{\boldsymbol{\beta}}_{\text{MLE}}),$$

which is a **biased estimate** of the true variance, σ^2. By contrast, the OLS estimator for this parameter is

$$\widehat{\sigma}^2_{\text{OLS}} := \frac{1}{n - p^*} \operatorname{RSS}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) = \frac{n}{n - p^*} \sigma^2_{\text{MLE}},$$

which can be shown to be **unbiased**. In practice, we tend to favor the OLS estimator, as the MLE for σ^2 **under-estimates** the variance of the residuals, which can lead to spurious statistical inference on the $\widehat{\beta}_j$'s.

2.4 Uniqueness of OLS Estimators

One can easily verify that $\widehat{\boldsymbol{\beta}}$ is a unique minimizer of $\operatorname{RSS}(\boldsymbol{\beta})$. In this section, we simply denote the OLS estimator by $\widehat{\boldsymbol{\beta}}$. It was shown that this quantity is a **critical point** (stationary point), since $\partial / \partial \boldsymbol{\beta} \operatorname{RSS}(\boldsymbol{\beta}) = 0$, when evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. However, such a critical point can be either a local minimizer, maximizer or a saddle point in $\mathbb{R}^{p^*}$. This critical point, however, can be shown to be a local minimizer, which is both global and unique:

i. $\widehat{\boldsymbol{\beta}}$ is a **local** minimizer, since $\partial^2 / \partial \boldsymbol{\beta}^2 \operatorname{RSS}(\boldsymbol{\beta})$ is **positive definite**, when evaluated at $\widehat{\boldsymbol{\beta}}$;

ii. $\widehat{\boldsymbol{\beta}}$ is a **global** minimizer, since $\operatorname{RSS}(\boldsymbol{\beta})$ is a **convex** function of $\boldsymbol{\beta}$ on $\mathbb{R}^{p^*}$;

iii. $\widehat{\boldsymbol{\beta}}$ is a **unique global** minimizer, since $\text{RSS}(\boldsymbol{\beta})$ is **strictly convex** on $\mathbb{R}^{p^*}$.

For the case of *real-valued* function such as $\text{RSS}(\boldsymbol{\beta})$, taking a vector as an argument, the *second partial derivative test* is conducted by computing the **Hessian** of $\text{RSS}(\boldsymbol{\beta})$. For any function $f : \mathbb{R}^m \mapsto \mathbb{R}$, with argument $\mathbf{x} \in \mathbb{R}^m$, the Hessian is the matrix of second derivatives defined by

$$H(f) := \frac{\partial^2}{\partial \mathbf{x} \partial \mathbf{x}} f(\mathbf{x}) = \begin{vmatrix} \frac{\partial^2}{\partial x_1^2} f(\mathbf{x}) & \frac{\partial^2}{\partial x_1 \partial x_2} f(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_1 \partial x_m} f(\mathbf{x}) \\ \frac{\partial^2}{\partial x_2 \partial x_1} f(\mathbf{x}) & \frac{\partial^2}{\partial x_2^2} f(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_2 \partial x_m} f(\mathbf{x}) \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_m \partial x_1} f(\mathbf{x}) & \frac{\partial^2}{\partial x_m \partial x_2} f(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_m^2} f(\mathbf{x}) \end{vmatrix}.$$

The Hessian of $\text{RSS}(\boldsymbol{\beta})$ can be computed as follows,

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \Big(2(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{y} \Big) = 2(\mathbf{X}^T\mathbf{X}),$$

Since $\mathbf{X}$ is **full-rank**, it follows that $\mathbf{X}^T\mathbf{X}$ is the *Gram matrix of a set of p^* linearly independent vectors*, which is a sufficient condition for $\mathbf{X}^T\mathbf{X}$ to be positive definite. That is, for any non-zero $\mathbf{a} \in \mathbb{R}^{p^*}$, we have

$$\mathbf{a}^T(\mathbf{X}^T\mathbf{X})\mathbf{a} > 0,$$

and therefore, the function $\text{RSS}(\boldsymbol{\beta})$ is **strictly convex** on $\mathbb{R}^{p^*}$, such that for any $t \in [0, 1]$, and any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{p^*}$, we have

$$\text{RSS}(t\boldsymbol{\beta}_1 + (1-t)\boldsymbol{\beta}_2) < t\,\text{RSS}(\boldsymbol{\beta}_1) + (1-t)\,\text{RSS}(\boldsymbol{\beta}_2).$$

In particular, $\partial^2/\partial\boldsymbol{\beta}^2\,\text{RSS}(\boldsymbol{\beta})$ is positive definite, when evaluated at $\widehat{\boldsymbol{\beta}}$, which implies that $\widehat{\boldsymbol{\beta}}$ is a local minimizer. The positive definiteness of $H(\text{RSS})$ is equivalent to the fact that all the **eigenvalues** of $H(\text{RSS})$ are positive. Moreover, this function is also **strictly convex**, and therefore any local minimizer is also a **unique global** minimizer. The latter condition, however, is strictly speaking not required, since $\widehat{\boldsymbol{\beta}}$ was already obtained in closed-form as the unique solution to $\partial/\partial\boldsymbol{\beta}\,\text{RSS}(\boldsymbol{\beta}) = 0$. Nonetheless, the second derivative test was required to identify whether $\widehat{\boldsymbol{\beta}}$ is a local minimum, maximum or a saddle point.

2.5 MLE Estimators and Fisher Information Matrix

We can similarly verify that the MLE for multiple regression under normal errors is also a unique global maximizer. The Hessian of the criterion used to compute the MLE is given by the matrix of second partial derivatives of the log-likelihood,

$$H(\log L) := \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}; \mathbf{y}, \mathbf{X}).$$

Using the derivations in equation (3), we obtain

$$H(\log L) := -\frac{1}{\sigma^2}(\mathbf{X}^T\mathbf{X}),$$

which is necessarily **negative definite**, since $\mathbf{X}^T\mathbf{X} > 0$. Thus, the MLE for this model is a unique global maximizer. Observe, however, that the uniqueness of $\widehat{\boldsymbol{\beta}}_{\text{MLE}}$ entirely depends on the fact that $\mathbf{X}$ is a **full-rank** matrix, as otherwise we would not be able to invert $\mathbf{X}^T\mathbf{X}$.

The Hessian matrix of the log-likelihood is of special interest in statistics, because it can be given a substantive geometrical interpretation dating back to the work of Ronald Fisher. We define the (observed) **Fisher information matrix** with respect to the full vector of parameters $\boldsymbol{\theta} := (\boldsymbol{\beta}, \sigma^2)$,

$$\mathcal{I}(\boldsymbol{\theta}) := -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \log L(\boldsymbol{\beta}, \sigma_2; \mathbf{y}, \mathbf{X}).$$

This quantity –when it exists, i.e. when the log-likelihood is twice-differentiable– defines a *Riemannian metric* on the manifold corresponding to the parameter space of the log-likelihood. For the Normal distribution, a special type of parameterization leads to a **diagonal** Fisher information matrix, thereby indicating that the mean and the variance of this distribution forms a set of orthogonal coordinates.

References

Casella, G. and Berger, R. (2002). *Statistical Inference (2nd edition)*. Duxbury, New York.

Christensen, R. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models (4th edition)*. Springer, New York.