

1 Consistency of the MLE

If a random variable or vector X has a (univariate or multivariate) density $f(x|\theta)$ or a pmf $p(x|\theta)$, then $\ell(\theta|X)$ denotes the log likelihood. In the density case, we have $\ell(\theta|X) = \log f(X|\theta)$. Note the use of capital letters, to emphasize that we are dealing with *random variables* here. Below we assume that the parameter θ belongs to (some subset of) \mathbb{R} and that the partial derivatives of $\ell(\theta|X)$ exist. We denote $\dot{\ell}(\theta|X) = \frac{\partial}{\partial\theta}\ell(\theta|X)$; likewise $\ddot{\ell}(\theta|X) = \frac{\partial^2}{\partial\theta^2}\ell(\theta|X)$.

If $X = (X_1, \dots, X_n)$ is a sample (independent random variables having the same distribution, they are iid) with marginal densities $f(x_i|\theta)$, then $\ell(\theta|X) = \sum_{i=1}^n \ell(\theta|X_i)$.

Let θ_0 denotes the ‘true’ (the one you want to know by estimation) and θ an arbitrary parameter value. The MLE when one observes a sample, can be found by maximizing

$$\frac{1}{n} \sum_{i=1}^n \ell(\theta|X_i),$$

an average of iid random variables. By the LLN, this quantity converges in \mathbb{P}_{θ_0} -probability to their common expectation $\mathbb{E}_{\theta_0} \ell(\theta|X_1) =: g(\theta)$. It is then reasonable to think that the MLE $\hat{\theta}$ converges to the maximum of $\theta \mapsto g(\theta)$. We show that the latter has a maximum at $\theta = \theta_0$. The first order condition is that $\dot{g}(\theta_0) = 0$ which we check as follows. First we compute

$$\dot{\ell}(\theta|X_1) = \frac{\partial}{\partial\theta} \log f(X_1|\theta) = \frac{\dot{f}(X_1|\theta)}{f(X_1|\theta)}. \quad (1.1)$$

Interchanging differentiation and integration (expectation), we have

$$\begin{aligned} \dot{g}(\theta) &= \mathbb{E}_{\theta_0} \dot{\ell}(\theta|X_1) \\ &= \int \frac{\dot{f}(x|\theta)}{f(x|\theta)} f(x|\theta_0) dx. \end{aligned}$$

Hence $\dot{g}(\theta_0) = \int \dot{f}(x|\theta_0) dx$. Interchanging differentiation and integration again, we obtain $\dot{g}(\theta_0) = \frac{\partial}{\partial\theta_0} \int f(x|\theta_0) dx$. This is equal to zero, since the integral equals one. We conclude $\dot{g}(\theta_0) = 0$. To know that θ_0 is a maximum, one has to verify that $\ddot{g}(\theta_0) < 0$. This can be done along the same lines, as you should verify, see also Remark 2.3 below. Then we hope that θ_0 is the only local maximum and thus a global maximum. This is actually true, but needs another argument and an extra condition.

The rough idea is thus that for large n the MLE should be ‘close’ to the maximizer of $\theta \mapsto g(\theta)$, which is shown to be θ_0 . Additional mathematics is needed to justify this rough idea and to conclude that indeed consistency of the MLE $\hat{\theta}_n$ holds, when the sample size n tends to infinity: $\hat{\theta}_n \xrightarrow{\mathbb{P}_{\theta_0}} \theta_0$, whatever the value of θ_0 .

2 Fisher information

We define the Fisher information and derive some properties. The importance of the Fisher information is explained in the next section. Below X is some random variable, or a random vector or a sample, depending on the context. We assume that all derivatives that we encounter exist.

Definition 2.1 Let X have a distribution depending on a parameter θ . The Fisher information about θ contained in X , denoted $I(\theta|X)$, is defined by $\mathbb{E}_\theta(\dot{\ell}(\theta|X)^2)$, usually simply called Fisher information. Note that $I(\theta|X) \geq 0$.

Proposition 2.2 Under some regularity conditions we have

1. It holds that $\mathbb{E}_\theta \dot{\ell}(\theta|X) = 0$.
2. The Fisher information also satisfies $I(\theta|X) = \text{Var}_\theta \dot{\ell}(\theta|X)$.
3. An alternative formula is $I(\theta|X) = -\mathbb{E}_\theta \ddot{\ell}(\theta|X)$.
4. For a sample $X = (X_1, \dots, X_n)$ we have

$$I(\theta|X) = \sum_{i=1}^n I(\theta|X_i) = nI(\theta|X_1).$$

In this case we usually write $I(\theta)$ instead of $I(\theta|X_1)$ and we have thus $I(\theta|X) = nI(\theta)$.

Proof We give the proof of the first two items for the case where X is a random variable with a density $f(x|\theta)$. If X is higher dimensional you only need more integrals.

1. Recall (1.1) and use X instead of X_1 . Then

$$\mathbb{E}_\theta \dot{\ell}(\theta|X) = \int \frac{\dot{f}(x|\theta)}{f(x|\theta)} f(x|\theta) dx = \int \dot{f}(x|\theta) dx = \frac{\partial}{\partial \theta} \int f(x|\theta) dx = 0.$$

Note that we interchanged integration and expectation.

2. Recall that for any random variable Y with expectation zero, one has that $\text{Var} Y = \mathbb{E}Y^2$ and use the previous assertion with $Y = \dot{\ell}(\theta|X)$.
3. Start with the result of the first assertion, which reads in integral form $0 = \int \dot{\ell}(\theta|x) f(x|\theta) dx$. Differentiate under the integral sign, use the product rule and (1.1) to get

$$\begin{aligned} 0 &= \int (\ddot{\ell}(\theta|x) f(x|\theta) + \dot{\ell}(x|\theta) \dot{f}(x|\theta)) dx \\ &= \int (\ddot{\ell}(\theta|x) f(x|\theta) + \dot{\ell}(x|\theta)^2 f(x|\theta)) dx \\ &= \mathbb{E}_\theta \ddot{\ell}(\theta|X) + \mathbb{E}_\theta \dot{\ell}(X|\theta)^2 \\ &= \mathbb{E}_\theta \ddot{\ell}(\theta|X) + I(\theta|X), \end{aligned}$$

by definition of $I(\theta|X)$. The results follows.

4. In case we are dealing with a sample we have with $x = (x_1, \dots, x_n)$ the product rule for the multivariate (joint) density $f(x|\theta) = \prod_{i=1}^n f(x_i|\theta)$. Hence, by taking logarithms, replacing x by X , one obtains $\ell(\theta|X) = \sum_{i=1}^n \ell(\theta|X_i)$ and then by differentiation $\dot{\ell}(\theta|X) = \sum_{i=1}^n \dot{\ell}(\theta|X_i)$. Note that we now have a sum of independent random variables and the sum rule for the variance applies: $\text{Var}_\theta \dot{\ell}(\theta|X) = \sum_{i=1}^n \text{Var}_\theta \dot{\ell}(\theta|X_i)$. Knowing the second assertion, we get $I(\theta|X) = \sum_{i=1}^n I(\theta|X_i)$. But since the X_i all have the same distribution, all $I(\theta|X_i)$ are equal to $I(\theta|X_1)$, which completes the proof. □

Remark 2.3 The function g in the previous section has the property that $\ddot{g}(\theta_0) = -I(\theta_0|X_1) \leq 0$. Show that the equality holds true and conclude that g has a (local) maximum in θ_0 .

3 Asymptotics for the MLE

The importance of the Fisher information is mainly because of the following theorem. Recall the notation of the previous section.

Theorem 3.1 *Under regularity conditions (for instance, differentiation of the likelihood w.r.t. θ is possible, etc.) one has the following central limit theorem type of result for the distribution of the MLE. Let $\hat{\theta}_n$ be the MLE based on a sample of n observations. Then*

$$\sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{d_\theta} Z,$$

where the distribution of Z is standard normal.

Remark 3.2 Two remarks. The convergence in distribution takes place under the condition that the distribution of $\hat{\theta}_n$ is used with the parameter value θ , just as consistency also involves this parameter, when we write $\hat{\theta}_n \xrightarrow{\mathbb{P}_\theta} \theta$. This explains why we use the symbol $\xrightarrow{d_\theta}$ instead of \xrightarrow{d} , as we did when discussing convergence in distribution. The second remark applies to the (exceptional) situation in which $\sqrt{nI(\theta)}(\hat{\theta}_n - \theta)$ would exactly have a $N(0, 1)$ distribution. Then we would have $\mathbb{E}_\theta(\sqrt{nI(\theta)}(\hat{\theta}_n - \theta)) = 0$ and $\text{Var}_\theta(\sqrt{nI(\theta)}(\hat{\theta}_n - \theta)) = 1$, which is equivalent to $\mathbb{E}_\theta \hat{\theta}_n = \theta$ and $\text{Var}_\theta(\hat{\theta}_n) = \frac{1}{nI(\theta)}$, as you easily check. Realizing that these properties only hold in a certain asymptotic sense (which has to be treated with care!), we paraphrase them by saying that the MLE is asymptotically unbiased and has an asymptotic variance equal to $\frac{1}{nI(\theta)}$.

Sketch of the proof We have for an arbitrary differentiable function f the Taylor expansion $f(y) = f(x) + (y-x)f'(x) + \dots$. Apply this to $f(\cdot) = \ell(\cdot|X)$, $y = \hat{\theta}_n$ and $x = \theta$ and use $\dot{\ell}(\hat{\theta}_n|X) = 0$ to get

$$0 = \dot{\ell}(\theta|X) + (\hat{\theta}_n - \theta)\ddot{\ell}(\theta|X) + \dots,$$

where we neglect the higher order remainder terms since $\hat{\theta}_n - \theta$ is small for large n by consistency of the MLE. It follows that we have the approximation

$$\hat{\theta}_n - \theta \approx -\frac{\dot{\ell}(\theta|X)}{\ddot{\ell}(\theta|X)}$$

and therefore, use a bit of elementary algebra,

$$\sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \approx \frac{\frac{1}{\sqrt{nI(\theta)}}\dot{\ell}(\theta|X)}{-\frac{1}{nI(\theta)}\ddot{\ell}(\theta|X)}.$$

We treat numerator and denominator separately. Let

$$Z_i = \frac{\dot{\ell}(\theta|X_i)}{\sqrt{I(\theta)}}.$$

Then we have $\mathbb{E}_\theta Z_i = 0$ and $\text{Var}_\theta Z_i = 1$ by Proposition 2.2. Moreover the Z_i are iid. The numerator we can thus rewrite as $N_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i$ to which we apply the Central Limit Theorem. It converges in distribution to a standard normal random variable Z , $N_n \xrightarrow{d_\theta} Z$.

To treat the denominator, call it D_n , we put

$$W_i = -\frac{\ddot{\ell}(\theta|X_i)}{I(\theta)}.$$

Invoking Proposition 2.2 again, we see that $\mathbb{E}_\theta W_i = 1$. Hence $D_n = \frac{1}{n} \sum_{i=1}^n W_i$. By the Law of large numbers (the W_i are iid), D_n converges in probability to the common expectation of the W_i and we obtain $D_n \xrightarrow{\mathbb{P}_\theta} 1$.

Combining the results for N_n and D_n , we find

$$\frac{N_n}{D_n} \xrightarrow{d_\theta} Z,$$

by the rules (see the slides) for combining convergence in probability and convergence in distribution. \square

4 Asymptotic optimality of the MLE

First we treat the Cramér-Rao bound on the variance of an unbiased estimator.

Theorem 4.1 (Cramér-Rao) *Let $\hat{\theta} = \hat{\theta}(X)$ be an unbiased estimator of θ , computed from a random vector X . Let $I(\theta|X)$ be the Fisher information. Then the mean squared error of $\hat{\theta}$, which is in this case equal to its variance, satisfies*

$$\text{Var}_\theta \hat{\theta} \geq \frac{1}{I(\theta|X)}.$$

In particular, when $X = (X_1, \dots, X_n)$ is a sample, then we have $\text{Var}_\theta \hat{\theta} \geq \frac{1}{nI(\theta)}$.

Proof Recall that the correlation coefficient $\rho = \rho(Y, Z)$ of a pair of random variables always lies between -1 and $+1$, so $0 \leq \rho^2 \leq 1$. This implies that always

$$\text{Cov}(Y, Z)^2 \leq \text{Var}(Y)\text{Var}(Z).$$

We choose $Y = \dot{\ell}(\theta|X)$, $Z = \hat{\theta}(X)$ and compute for these the variances and covariance. We know from Proposition 2.2 that $\text{Var}_\theta(\dot{\ell}(\theta|X)) = I(\theta|X)$. We are interested in $\text{Var}_\theta(\hat{\theta}(X))$ and so the only thing left to compute is the covariance $\text{Cov}_\theta(\dot{\ell}(\theta|X), \hat{\theta}(X))$. Since $\mathbb{E}_\theta \dot{\ell}(\theta|X) = 0$, we have $\text{Cov}_\theta(\dot{\ell}(\theta|X), \hat{\theta}(X)) = \mathbb{E}_\theta(\dot{\ell}(\theta|X)\hat{\theta}(X))$. We compute this expectation as an integral (under the temporary assumption that X is real valued). In the one but last equation below we use that $\hat{\theta}(X)$ is unbiased, $\mathbb{E}_\theta \hat{\theta}(X) = \theta$.

$$\begin{aligned} \mathbb{E}_\theta(\dot{\ell}(\theta|X)\hat{\theta}(X)) &= \int \dot{\ell}(\theta|x)\hat{\theta}(x)f(x|\theta) \, dx \\ &= \int \frac{\dot{f}(x|\theta)}{f(x|\theta)}\hat{\theta}(x)f(x|\theta) \, dx \\ &= \int \dot{f}(x|\theta)\hat{\theta}(x) \, dx \\ &= \frac{\partial}{\partial \theta} \int f(x|\theta)\hat{\theta}(x) \, dx \\ &= \frac{\partial}{\partial \theta} \theta \\ &= 1. \end{aligned}$$

Having computed all relevant quantities, we deduce

$$1 \leq I(\theta|X)\text{Var}_\theta \hat{\theta}(X),$$

from which the result follows. \square

The content of Theorem 4.1 is that no unbiased estimator can have a variance (which is here equal to the MSE) smaller than $\frac{1}{I(\theta|X)}$, which can thus be considered as the best possible (best refers to minimum mean squared error for all θ). If X is a sample (X_1, \dots, X_n) , the lower bound on the variance in Theorem 4.1 becomes $\frac{1}{nI(\theta)}$. Where have we seen this quantity before? Indeed, in Theorem 3.1 on the asymptotic normality of the MLE, and the discussion after this theorem in Remark 3.2. There we have argued that, from a certain asymptotic point of view, the MLE is almost unbiased for large n with asymptotic variance approximately equal to $\frac{1}{nI(\theta)}$. Hence the MLE achieves for large n , asymptotically the lowest possible value in the Cramér-Rao theorem. This phenomenon can be summarized by saying that the MLE is asymptotically optimal for the mean squared error criterion.

5 Results on regression

First a summary of the model and the estimators. The model is

$$Y = X\beta + e, \quad (5.2)$$

with e n -dimensional, $\beta \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$. Moreover, it is assumed that X has full rank (equal to p) and e has a multivariate normal distribution with mean vector zero and covariance matrix equal to $\sigma^2 I_n$. The LS estimator of β is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

and satisfies

$$\hat{\beta} = \beta + (X^\top X)^{-1} X^\top e.$$

Therefore $\hat{\beta}$ has a multivariate normal distribution, is unbiased and $\text{Cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$. The vector of residuals is

$$\hat{e} = Y - X\hat{\beta} = (I - X(X^\top X)^{-1} X^\top)Y = (I - X(X^\top X)^{-1} X^\top)e.$$

Let $Q = I - X(X^\top X)^{-1} X^\top$. Verify that Q is idempotent, $Q^2 = Q$. The residual sum of squares is defined as $\hat{e}^\top \hat{e}$ and thus equal to $e^\top Qe$.

Theorem 5.1 *The following properties hold.*

- (i) $\hat{\beta}$ and $\hat{e}^\top \hat{e}$ are independent.
- (ii) $\frac{1}{\sigma^2} \hat{e}^\top \hat{e}$ has a χ_{n-p}^2 distribution.

Proof (i) Consider the matrix $V = (X^\top X)^{-1/2} X^\top$. Verify that $VV^\top = I_p$, from which we conclude that $V \in \mathbb{R}^{p \times n}$ has orthonormal rows. Take now a matrix $W \in \mathbb{R}^{(n-p) \times p}$ that has orthonormal rows as well and whose rows are also orthogonal to those of V . Then the matrix

$$U = \begin{pmatrix} V \\ W \end{pmatrix} \in \mathbb{R}^{n \times n}$$

is orthonormal too, so $UU^\top = I_n$ and $U^\top U = I_n$. Define $Z = Ue$. Then Z is multivariate normal, $\mathbb{E}Z = 0$, and has covariance matrix $\mathbb{E}(ZZ^\top) = \mathbb{E}(Uee^\top U^\top) = U\mathbb{E}(ee^\top)U^\top = U\sigma^2 I_n U^\top = \sigma^2 I_n$. We conclude that Z has independent components.

Split Z as

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} Ve \\ We \end{pmatrix}.$$

Then Z_1 and Z_2 are independent random vectors, both with having a multivariate normal distribution. Moreover, $\hat{\beta} = (X^\top X)^{-1/2} Z_1 + \beta$. Furthermore, we have

$$Z^\top Z = e^\top U^\top Ue = e^\top I_n e = e^\top e$$

and

$$Z^\top Z = Z_1^\top Z_1 + Z_2^\top Z_2 = e^\top X(X^\top X)^{-1}Xe + Z_2^\top Z_2.$$

Combining the two expressions for $Z^\top Z$, we find

$$Z_2^\top Z_2 = e^\top (I - X(X^\top X)^{-1}X)e = e^\top Qe$$

and therefore $\hat{e}^\top \hat{e} = Z_2^\top Z_2$. Since $\hat{\beta}$ is a function of Z_1 and $\hat{e}^\top \hat{e}$ is a function of Z_2 , we have proved the first assertion.

(ii) The vector Z_2 has independent elements, all having a $N(0, \sigma^2)$ distribution. Since it has $n - p$ elements, $\frac{1}{\sigma^2} Z_2^\top Z_2$ has a χ_{n-p}^2 distribution. Because $Z_2^\top Z_2 = \hat{e}^\top \hat{e}$, the second assertion now follows. \square

It follows Theorem 5.1 that $\mathbb{E}(\hat{e}^\top \hat{e}) = \sigma^2(n - p)$. Hence

$$\widehat{\sigma^2} := \frac{\hat{e}^\top \hat{e}}{n - p}$$

is an unbiased estimator of σ^2 . If $\sigma_{\hat{\beta}_i}^2$ denotes the variance of $\hat{\beta}_i$, then

$$\sigma_{\hat{\beta}_i}^2 = \sigma^2 (X^\top X)^{-1}_{ii},$$

which has

$$s_{\hat{\beta}_i}^2 := \widehat{\sigma^2} (X^\top X)^{-1}_{ii}$$

as an unbiased estimator.

Corollary 5.2 *Let $i \in \{0, \dots, p - 1\}$. The statistic*

$$T_i := \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}}$$

has a t_{n-p} distribution.

Proof We know that $\hat{\beta}_i$ has a $N(\beta_i, \sigma^2 (X^\top X)^{-1}_{ii})$ distribution. Therefore

$$Z = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 (X^\top X)^{-1}_{ii}}}$$

has a standard normal distribution and

$$\hat{\beta}_i - \beta_i = Z \sqrt{\sigma^2 (X^\top X)^{-1}_{ii}}.$$

We also know from Theorem 5.1 that $R = \frac{\hat{e}^\top \hat{e}}{\sigma^2}$ has a χ_{n-p}^2 distribution. Then we can write

$$s_{\hat{\beta}_i}^2 = \sigma^2 \frac{R}{n - p} (X^\top X)^{-1}_{ii}.$$

It now follows that

$$T_i = \frac{Z \sqrt{\sigma^2 (X^\top X)^{-1}_{ii}}}{\sqrt{\sigma^2 \frac{R}{n-p} (X^\top X)^{-1}_{ii}}} = \frac{Z}{\sqrt{R/(n-p)}},$$

which is the quotient of a standard normal random variable and the root of a random variable that has a χ^2 distribution divided by its number of degrees of freedom. By Theorem 5.1 it is also a ratio of independent random variables and the result follows by the definition of the t -distribution. \square