TI Statistics - Extra notes

P.J.C. Spreij

this version: August 24, 2020

1 Probability from a measure theoretic perspective

Although it will play no role in this course, for completeness we present an outline of the basic of foundations probability theory from a measure theoretic perspective. The central object is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which we will explain now.

Think of a probabilistic experiment on the background for which we have Ω as the set of all possible outcomes. An *event* is usually understood as a subset of Ω. But we will require more, the set of all subsets of Ω that deserve to be called events, denoted $\mathcal{F}$, has to obey certain requirements, it has to be a σ-algebra. This means that $\mathcal{F}$ is such that

- $\emptyset \in \mathcal{F}$,
- If $A \in \mathcal{F}$, then also its complement A^c is an element of $\mathcal{F}$,
- If $A_1, A_2, \ldots$ is a sequence of sets in $\mathcal{F}$, then also the union $\bigcup_{i=1}^{\infty} A_i$ belongs to $\mathcal{F}$.

It is a nice exercise to show that also finite unions like $A_1 \cup A_2$ belong to $\mathcal{F}$, whenever $A_1, A_2 \in \mathcal{F}$. Also finite and countable intersections $A_1 \cap A_2$ and $\bigcap_{i=1}^{\infty} A_i$ belong to $\mathcal{F}$, if the A_i belong to it. It follows that from certain events one can construct new events by taking intersections, unions, and complements, even countably infinite often. So all set theoretic operations applied to events yield events again, as long as they are performed at most countably often.

If the set Ω is finite or countable, one usually take the power set of Ω (all its subsets) as the collection of events $\mathcal{F}$. But if Ω is countably infinite, like $\Omega = \mathbb{R}$ or $\Omega = (0, 1)$, for technical reasons one takes a smaller collection. In the latter two examples, one usually takes the *Borel sets* (denoted $\mathcal{B}$), these are the sets that can be generated by at most countably often applied set theoretic operations to all open intervals. For example, if $\Omega = \mathbb{R}$, then by definition an interval $(-\infty, a)$ is an element of $\mathcal{B}$, but then also $[a, \infty)$. Also every singleton belongs to $\mathcal{B}$, since $\{a\} = \cap_{n=1}^{\infty}(a - 1/n, a + 1/n)$. Other sets in $\mathcal{B}$ are $(-\infty, a]$, $(a, b]$, $[a, b)$, $\mathbb{Q}$ (the set of rational numbers), etc. In fact any '*normal*' subset of $\mathbb{R}$ will be in $\mathcal{B}$, although formally there is a tautology here (all sets in $\mathcal{B}$ are 'normal'...).

The notation for the probability of a set $A \in \mathcal{F}$ is $\mathbb{P}(A)$, which resembles the notation $f(x)$ in case one deals with a function f. Indeed, a probability $\mathbb{P}$, also known as a *probability measure*, is a function too, defined on the collection of events $\mathcal{F}$. More precisely, we say that a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability (measure), if

- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$,
- for *disjoint* events $A_i \in \mathcal{F}$ it holds that $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Note that by the property of $\mathcal{F}$ being a σ-algebra, automatically the union above is in $\mathcal{F}$ and so its probability is defined. Note also that for disjoint A_1 and A_2, both in $\mathcal{F}$, we have the familiar rule $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$ (you check!). Other rules for events are $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$ and $\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i)$ if $\cup_{i=1}^{\infty} B_i = \Omega$ (cutting A up in slices B_i).

A random variable X is by definition a function on Ω, so $X : \Omega \to \mathbb{R}$, that is required to be *measurable*, i.e. $\{X \in B\} \in \mathcal{F}$ for every Borel set B, where $\{X \in B\}$ is shorthand notation for $\{\omega \in \Omega : X(\omega) \in B\}$. For example, every set $\{X \leq x\}$ is an element of $\mathcal{F}$ for a random variable X (here you take $B = (-\infty, x]$, which indeed belongs to $\mathcal{B}$). In fact, it is possible to show that if all sets $\{X \leq x\}$ ($x \in \mathbb{R}$) are elements of $\mathcal{F}$, then X is measurable.

The consequence is that for random variables X now automatically the probabilities $\mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$ are well defined. Moreover, we also have the rule $\mathbb{P}(X \in B_1 \cup B_2) = \mathbb{P}(X \in B_1) + \mathbb{P}(X \in B_2)$ for disjoint B_1, B_2 in $\mathcal{B}$. The probabilities $F(x) := \mathbb{P}(X \leq x)$ are the values of a function $F : \mathbb{R} \to [0, 1]$, called the *distribution function* of X. Here is a nice exercise: show that F is non-decreasing and right-continuous. It is a theorem that the function F uniquely fixes all probabilities $\mathbb{P}(X \in B)$ for $B \in \mathcal{B}$. The collection of these probabilities form the *distribution* of X. Moreover the mapping $\mathbb{P}^X : \mathcal{B} \to [0, 1]$ defined by $\mathbb{P}^X(B) := \mathbb{P}(X \in B)$ is a probability measure (in the above sense) on $\mathcal{B}$.

Random *vectors* X will be considered as vectors of random variables X_i. A two-dimensional random vector is sometimes denoted as a row (X_1, X_2) or as a column $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, depending on the circumstances, or on what is notationally more convenient.

As said above, this measure theoretic set up will play no role in this course, but you should have seen once in your life, that also probability theory is based on definitions and axioms (of which we have only presented the most basic ones), like any other branch of mathematics.

2 Some facts on stochastic convergence

Inn the next propositions we collect some facts on stochastic convergence. In particular we first present how the different modes of convergence are related and how convergence is preserved under transformations.

Proposition 2.1 *Let $X, X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be random variables, c a real constant.*

1. *If $X_n \xrightarrow{P} X$, then also $X_n \xrightarrow{d} X$.*
2. *If $X_n \xrightarrow{d} c$, then also $X_n \xrightarrow{P} c$.*
3. *If $X_n \xrightarrow{P} c$, then also $g(X_n) \xrightarrow{P} g(c)$, if g is a continuous at c. Similar statement for $\xrightarrow{d}$.*

4. *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{d} c$, *then* $g(X_n, Y_n) \xrightarrow{d} g(X, c)$, *if* g *is a continuous function (on* $\mathbb{R}^2$*).*

In the next proposition we collect some 'calculus rules', that give rules for combining results, when one deals with two 'convergent' sequences, even if the modes of convergence differ.

Proposition 2.2 *Let* $X, X_1, X_2, \ldots$ *and* $Y_1, Y_2, \ldots$ *be random variables,* c *a real constant.*

1. *If* $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$, *then also* $X_n \pm Y_n \xrightarrow{P} X \pm Y$.
2. *If* $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$, *then also* $X_n Y_n \xrightarrow{P} XY$, *and also* $X_n/Y_n \xrightarrow{P} X/Y$ *provided* $P(Y \neq 0) = 1$.
3. *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{P} c$, *then also* $X_n \pm Y_n \xrightarrow{d} X \pm c$.
4. *If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{P} c$, *then also* $X_n Y_n \xrightarrow{d} Xc$, *and* $X_n/Y_n \xrightarrow{d} X/c$ *provided* $c \neq 0$.

3 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 $\dot{\ell}(\theta|X) = \frac{\partial}{\partial\theta}\ell(\theta|X)$; the notation $\ddot{\ell}(\theta|X)$ should now be obvious.

If $X = (X_1, \ldots, X_n)$ 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 $\mathbb{P}_{\theta_0}$-probability to their common expectation $\mathbb{E}_{\theta_0}\ell(\theta|X_1) =: g(\theta)$. It is then reasonably 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

$$\dot{\ell}(\theta|X_1) = \frac{\partial}{\partial\theta}\log f(X_1|\theta) = \frac{\dot{f}(X_1|\theta)}{f(X_1|\theta)}. \tag{3.1}$$

Interchanging differentiation and integration (expectation), we have

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

Hence $\dot{g}(\theta_0) = \int \dot{f}(x|\theta_0)\,\mathrm{d}x$. Interchanging differentiation and integration again, we obtain $\dot{g}(\theta_0) = \frac{\partial}{\partial\theta_0}\int f(x|\theta_0)\,\mathrm{d}x$. 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 4.3 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: $\hat{\theta}_n \overset{\mathbb{P}_{\theta_0}}{\to} \theta_0$, whatever the value of θ_0.

4 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 4.1 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 $\mathbb{E}_\theta\big(\dot{\ell}(\theta|X)^2\big)$, usually simply called Fisher information. Note that $I(\theta|X) \geq 0$.

Proposition 4.2 *Under some regularity conditions we have*

(i) It holds that $\mathbb{E}_\theta\dot{\ell}(\theta|X) = 0$.

(ii) The Fisher information also satisfies $I(\theta|X) = \mathbb{V}\mathrm{ar}_\theta\,\dot{\ell}(\theta|X)$.

(iii) An alternative formula is $I(\theta|X) = -\mathbb{E}_\theta\ddot{\ell}(\theta|X)$.

(iv) For a sample $X = (X_1, \ldots, X_n)$ *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 have* $I(\theta|X) = nI(\theta)$.

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.

(i) Recall (3.1) and use X instead of X_1. Then

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

Note that we interchanged integration and expectation.

(ii) Recall that for any random variable Y with expectation zero, one has that $\mathbb{V}\mathrm{ar}\, Y = \mathbb{E}Y^2$ and use the previous assertion with $Y = \dot\ell(\theta|X)$.

(iii) Start with the result of the first assertion, which reads in integral form $0 = \int \dot\ell(\theta|x) f(x|\theta)\, \mathrm{d}x$. Differentiate under the integral sign, use the product rule and (3.1) to get

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

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

(iv) In case we are dealing with a sample we have with $x = (x_1, \ldots, x_n)$ 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: $\mathbb{V}\mathrm{ar}_\theta\, \dot\ell(\theta|X) = \sum_{i=1}^n \mathbb{V}\mathrm{ar}_\theta\, \dot\ell(\theta|X_i)$. 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 4.3 The function g in the previous section has the property that $\ddot g(\theta_0) = -I(\theta_0|X_1) \le 0$. Show that the equality holds true and conclude that g has a (local) maximum in θ_0.

5 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 5.1 *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 5.2 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 $\hat{\theta}_n \stackrel{\mathbb{P}_\theta}{\to} \theta$. This explains why use the symbol $\stackrel{d_\theta}{\to}$ instead of $\stackrel{d}{\to}$, 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 $\mathbb{V}\mathrm{ar}_\theta\big(\sqrt{nI(\theta)}(\hat{\theta}_n - \theta)\big) = 1$, which is equivalent to $\mathbb{E}_\theta\, \hat{\theta}_n = \theta$ and $\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) = \frac{1}{nI(\theta)}$, 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 $f(y) = f(x) + (y-x)f'(x) + \cdots$. Apply this to $f(\cdot) = \dot{\ell}(\cdot|X)$, $y = \hat{\theta}_n$ and $x = \theta$ and use $\dot{\ell}(\hat{\theta}_n|X) = 0$ to get

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

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 $\mathbb{V}\mathrm{ar}_\theta\, Z_i = 1$ by Proposition 4.2. 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, $N_n \stackrel{d_\theta}{\to} Z$.

To treat the denominator, call it D_n, we put

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

Invoking Proposition 4.2 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 $D_n \stackrel{\mathbb{P}_\theta}{\to} 1$.

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. □

6 Asymptotic optimality of the MLE

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

Theorem 6.1 (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*

$$\mathbb{V}\mathrm{ar}_\theta\, \hat{\theta} \geq \frac{1}{I(\theta|X)}.$$

In particular, when $X = (X_1, \ldots, X_n)$ is a sample, then we have $\mathbb{V}\mathrm{ar}_\theta\, \hat{\theta} \geq \frac{1}{nI(\theta)}$.

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

$$\mathrm{Cov}(Y, Z)^2 \leq \mathbb{V}\mathrm{ar}(Y)\mathbb{V}\mathrm{ar}(Z).$$

We choose $Y = \dot{\ell}(\theta|X)$, $Z = \hat{\theta}(X)$ and compute for these the variances and covariance. We know from Proposition 4.2 that $\mathbb{V}\mathrm{ar}_\theta(\dot{\ell}(\theta|X)) = I(\theta|X)$. We are interested in $\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}(X))$ and so the only thing left to compute is the covariance $\mathrm{Cov}_\theta(\dot{\ell}(\theta|X), \hat{\theta}(X))$. Since $\mathbb{E}_\theta \dot{\ell}(\theta|X) = 0$, we have $\mathrm{Cov}_\theta(\dot{\ell}(\theta|X), \hat{\theta}(X)) = \mathbb{E}_\theta\big(\dot{\ell}(\theta|X)\hat{\theta}(X)\big)$. 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, $\mathbb{E}_\theta \hat{\theta}(X) = \theta$.

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

Having computed all relevant quantities, we deduce

$$1 \leq I(\theta|X)\mathbb{V}\mathrm{ar}_\theta\hat{\theta}(X),$$

from which the result follows. □

The content of Theorem 6.1 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 $(X_1, \ldots, X_n)$, the lower bound on the variance in Theorem 6.1 becomes $\frac{1}{nI(\theta)}$. Where have we seen this quantity before? Indeed, in Theorem 5.1 on the asymptotic normality of the MLE, and the discussion after this theorem in Remark 5.2. 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.

7 Results on multiple regression

Starting point is the multivariate regression model

$$Y = X\beta + \varepsilon,$$

where Y is an n-dimensional random (column) vector, X a $n \times p$ matrix, β the p-dimensional parameter vector and ε an n-dimensional random vector. We make the usual assumptions,

- the elements ε_i of the random vector ε are independent and have a common $N(0, \sigma^2)$ distribution,
- the matrix X has full rank equal to p (and so $X^\top X$ is invertible).

Note that ε has a multivariate normal distribution with mean vector zero and covariance matrix equal to $\sigma^2 I_n$, where I_n is the n-dimensional identity matrix. It then follows that Y has a multivariate normal distribution with mean vector $X\beta$ and covariance matrix equal to $\sigma^2 I_n$. The least squares estimator of β is denoted $\hat{\beta}$ and we have

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

The estimator of β_i is the i-th element of $\hat{\beta}$, denoted $\hat{\beta}_i$. One quickly shows that $\hat{\beta}$ is an unbiased estimator of β and

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

Moreover $\hat{\beta}$ has a multivariate normal distribution with mean β and covariance matrix $\sigma^2 (X^\top X)^{-1}$. It follows that $\mathbb{V}\mathrm{ar}\,\hat{\beta}_i = \sigma^2 (X^\top X)^{-1}_{ii}$.

Define $\hat{\varepsilon} = Y - X\hat{\beta}$. A simple computation shows that

$$\hat{\varepsilon} = Q\varepsilon,$$

where $Q = I_n - X(X^\top X)^{-1}X^\top$. Note that Q is symmetric, $Q^2 = Q$ and $QX = 0$ (verify this!). We will see below that

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

is an unbiased estimator of σ^2. This leads to a sensible estimator of $\mathbb{V}\mathrm{ar}\,\hat{\beta}_i$, $\hat{\sigma}^2(X^\top X)^{-1}_{ii}$. The main result of this section is the following

Theorem 7.1 *In the above regression set up the following hold true.*

(i) The random vectors $\hat{\beta}$ and the residuals $\hat{e}$ are independent.

(ii) The random variable $\frac{1}{\sigma^2}\hat{\varepsilon}^\top\hat{\varepsilon}$ has a χ^2-distribution with $n-p$ degrees on freedom, hence $\mathbb{E}\,\frac{\hat{\varepsilon}^\top\hat{\varepsilon}}{n-p} = \sigma^2$ and $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

(iii) The random variables ($i = 1, \ldots, p$)

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

have a t-distribution with $n-p$ degrees of freedom.

Proof We start with a series of preliminary facts, for which we need the *square root* R of the positive definite matrix $X^\top X$, the unique symmetric matrix that satisfies $R^2 = X^\top X$. Note that R is invertible (why?) and let $V = R^{-1}X^\top \in \mathbb{R}^{p\times n}$. We compute $VV^\top = R^{-1}X^\top XR^{-1} = I_p$ and see that the rows of V are orthonormal vectors. Hence there exists a matrix $W \in \mathbb{R}^{(n-p)\times n}$ whose rows are also orthonormal vectors such that the matrix

$$U = \begin{pmatrix} V \\ W \end{pmatrix}$$

is orthogonal of size $n \times n$, so $UU^\top = U^\top U = I_n$. It then follows that

$$I_n = \begin{pmatrix} V^\top & W^\top \end{pmatrix}\begin{pmatrix} V \\ W \end{pmatrix} = V^\top V + W^\top W.$$

Likewise it follows from

$$I_n = \begin{pmatrix} V \\ W \end{pmatrix}\begin{pmatrix} V^\top & W^\top \end{pmatrix} = \begin{pmatrix} VV^\top & VW^\top \\ WV^\top & WW^\top \end{pmatrix}$$

that $VV^\top = I_p$, $WW^\top = I_{n-p}$, $VW^\top = 0$.

Let

$$Z = U\varepsilon = \begin{pmatrix} V\varepsilon \\ W\varepsilon \end{pmatrix} =: \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}.$$

Then Z also has a multivariate normal distribution, with covariance matrix $\mathrm{Cov}(Z) = U\mathrm{Cov}(\varepsilon)U^\top = U(\sigma^2 I_n)U^\top = \sigma^2 UU^\top = \sigma^2 I_n$. We conclude, by an important property of the multivariate normal distribution, that Z has independent components, in particular Z_1 and Z_2 are independent. Note that Z_1 is p-dimensional and Z_2 is $(n-p)$-dimensional. All these considerations now pay off.

(i) First we have $\hat{\beta} = \beta + (X^\top X)^{-1} X\varepsilon = \beta + R^{-1}V\varepsilon = \beta + R^{-1}Z_1$, a linear transformation of Z_1. Second we have $Q = I_n - V^\top V = W^\top W$ and hence $\hat{\varepsilon} = Q\varepsilon = W^\top Z_2$, a linear transformation of Z_2. So, $\hat{\beta}$ and $\hat{\varepsilon}$ are independent.

(ii) Note that the $Z_2 = W\varepsilon$ above has a multivariate normal distribution with zero expectation vector and covariance matrix $\sigma^2 I_{n-p}$, as $WW^\top = I_{n-p}$. Hence the $n-p$ components of Z are independent and $\frac{1}{\sigma} Z_2$ has a multivariate standard normal $N(0, I_{n-p})$ distribution. It follows that $\frac{1}{\sigma^2} Z_2^\top Z_2$ has a χ^2_{n-p}-distribution. But from the proof of (i) we know that $\hat{\varepsilon}^\top \hat{\varepsilon} = Z_2^\top WW^\top Z_2 = Z_2^\top Z_2$, from which the assertion on the distribution follows. Since χ^2_{n-p} has expectation $n-p$, we find that $\mathbb{E}\frac{Z_2^\top Z_2}{\sigma^2} = n-p$ and hence $\mathbb{E}\,\hat{\sigma}^2 = \sigma^2$.

(iii) Let $\zeta_i = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{(X^\top X)^{-1}_{ii}}}$, $i = 1, \ldots, p$. A bit of rewriting yields

$$T_i = \frac{\zeta_i}{\sqrt{\frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{\sigma^2}/(n-p)}}.$$

Note the following three facts. The numerator and denominator are independent random variables (follows from (i)), the numerator has a standard normal distribution (why?), whereas in the denominator $\frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{\sigma^2}$ has a χ^2-distribution with $n-p$ degrees of freedom (as stated in (ii)), from which the result follows by the definition of a t-distribution. □

Remark 7.2 The case with $p = 1$ in the above theorem is a result that you have encountered earlier. Verify that this indeed the case, by inspecting how β and $\hat{\sigma}^2$ look like in this situation.