Probabilities and statistics
Lecturer: Prof. Dr. Sara van de Geer
Prof. Dr. Martin Larsson

Serie 11

May 18th, 2015

Q1. GAUSS-MARKOV THEOREM We want to study linear regression models. We do m experiments with explanatory variables $(x_i)_{i=1}^m \subseteq \mathbb{R}^n$ and with a scalar dependent variable $(y_i)_{i=1}^n \subseteq \mathbb{R}$. We suppose that for all i, the underlying model is given by

$$y_i = \beta \cdot x_i + \epsilon_i \quad \beta \in \mathbb{R}^n \tag{1}$$

where (ϵ_i) is a i.i.d sequence such that $\mathbb{E}(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. We want to estimate β. We say that $\tilde{\beta}$ is an unbiased estimator of β if

$$\mathbb{E}\left(\tilde{\beta}\right) = \beta.$$

Additionally we say that $\tilde{\beta}$ is linear if there exists a matrix, D, only depending on X such that $\tilde{\beta} = DY$ .We will also say that a matrix $A \lesssim B$ if $B - A$ is a positive semidefinite matrix.

(a) Show that (1) is equivalent to

$$Y = X\beta + \epsilon, \tag{2}$$

where $Y = \begin{pmatrix} y_1 \\ \vdots \\ y^t \end{pmatrix}$, $X = \begin{pmatrix} x_1^t \\ \vdots \\ x_m^t \end{pmatrix}$ and $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix}$.

(b) Show that the normal linear regression model (example 3.1 of the Skript) is a linear unbiased estimator. We will call its associated matrix K.

(c) Compute the covariance matrix of $\bar{\beta}$, the estimator of the normal linear regression model. **Hint:** Remember that if $Z \in \mathbb{R}^n$ is a random variable and C is a matrix then $V(CZ) = CZC^t$, where $V(\cdot)$ is the covariance matrix.

(d) Show that if $\tilde{\beta} = (K + C)Y$ is an unbiased estimator, then $CX = 0$.

(e) Show that the covariance matrix of $\tilde{\beta}$ is such that

$$V(\tilde{\beta}) \gtrsim V(\bar{\beta}).$$

Solution

(a) Just note that the coordinate i of (2) is given by

$$y_i = (X\beta)_i + \epsilon_i = \sum_{k=1}^{n} X_{ik}\beta_k + \epsilon_i = x_i \cdot \beta + \epsilon_i.$$

(b) We know that for the normal linear regression model is $\bar{\beta} := ((X^tX)^{-1}X)Y$, so it's a linear model. Let's compute its expected value

$$\begin{aligned}\mathbb{E}\left(\bar{\beta}\right) &= \mathbb{E}\left(((X^tX)^{-1}X^t)Y\right)\\ &= \mathbb{E}\left((X^tX)^{-1}X^t(X\beta+\epsilon)\right)\\ &= \beta + \mathbb{E}\left(\epsilon\right) = \beta,\end{aligned}$$

Then $\bar{\beta}$ is unbiased.

(c) We just have to compute

$$\begin{aligned}V(\bar{\beta}) &= V(((X^tX)^{-1}X^t)Y)\\ &= ((X^tX)^{-1}X^t)V(Y)((X^tX)^{-1}X^t)^t\\ &= \sigma^2(X^tX)^{-1}.\end{aligned}$$

(d) We just have to compute its expected value:

$$\begin{aligned}\mathbb{E}\left(\tilde{\beta}\right) &= \mathbb{E}\left(\bar{\beta}+CY\right)\\ &= \beta + C\mathbb{E}\left(X\beta+\epsilon\right)\\ &= (I+CX)\beta,\end{aligned}$$

given its expected value should be β for all $\beta \in \mathbb{R}^n$, then we have that $CX = 0$.

(e) We have to compute the covariance matrix of $\tilde{\beta}$

$$\begin{aligned}V(\tilde{\beta}) &= V(Cy) = CV(y)C^t = \sigma^2 CC^t\\ &= \sigma^2((X^tX)^{-1}X^t + D)(X(X^tX)^{-1} + D^t)\\ &= \sigma^2((X^tX)^{-1}X^tX(X^tX)^{-1} + (X^tX)^{-1}X^tD^t + DX(X^tX)^{-1} + DD^t)\\ &= \sigma^2(X^tX)^{-1} + \sigma^2(X^tX)^{-1}(\underbrace{DX}_{0})^t + \sigma^2\underbrace{DX}_{0}(X^tX)^{-1} + \sigma^2DD^t\\ &= \underbrace{\sigma^2(X^tX)^{-1}}_{V(\hat{\beta})} + \sigma^2DD'.\end{aligned}$$

To finish note that σ^2DD' is a positive semidefinitive matrix.

Q2. In a lake we want to estimate the amount of a certain type of fish. For this we mark 5 fishes and we let them mix with the others, when they are well mixed we fish 11, and we realize that there are 3 marked and 8 non-marked. What is the maximum-likelihood estimator for the amount of fishes?.

Solution: Define X the amount of marked fishes we fished. If there are N fishes in the lake, the probability of $X = 3$ is given by

$$\begin{aligned}\mathbb{P}_N(X=3) &= \frac{\binom{5}{3}\binom{N-5}{8}}{\binom{N}{11}}\mathbf{1}_{\{N\geq 13\}}\\ &= \frac{5!(N-5)!11!(N-11)!}{3!2!8!(N-13)!N!}\mathbf{1}_{\{N\geq 13\}} := g(N).\end{aligned}$$

We have to find $N_{\max} \in \mathbb{N}$ so that $g(N_{\max}) = \sup_{N\in\mathbb{N}} g(N)$. We have that for $N \geq 13$

$$\begin{aligned}\frac{g(N)}{g(N+1)} - 1 &= \frac{(N-12)(N+1)}{(N-4)(N-10)} - 1\\ &= \frac{3(N-17,\bar{3})}{(N-4)(N-10)},\end{aligned}$$

thus,

$$\frac{g(N)}{g(N+1)} \begin{cases} \leq 1 & \text{if } N \leq 17, \\ \geq 1 & \text{if } N \geq 18. \end{cases}$$

Then $N_{\max} = 18$.

Q3. Let $(X_i)_{i=1}^{2n+1}$ a sequence of i.i.d normal random variables with mean μ and variance σ unknown. We take two different estimators for μ:

$$\begin{aligned}T^{(1)}_{2n+1} &= \frac{1}{2n+1}\sum_{i=1}^{2n+1} X_i,\\ T^{(2)}_{2n+1} &= X_{(n+1)},\end{aligned}$$

where $X_{(1)} < X_{(2)} < ... < X_{(2n+1)}$ are the ordered results.

(a) With the help of the Central Limit Theorem find sequences $c^{(1)}_n$ and $c^{(2)}_n$ so that

$$\mathbb{P}\left(|T^{(i)}_{2n+1} - \mu| \leq c^{(i)}_n\right) \to 0.95.$$

(b) Find $q \in \mathbb{R}^+$ so that

$$\frac{c^2_{nq}}{c^1_n} \to 1,$$

how can we interpret, in words, q?.

Solution:

(a) We know that $T^{(1)}_{2n+1} \sim N\left(\mu, \frac{\sigma}{\sqrt{2n+1}}\right)$, then

$$\begin{aligned}&\mathbb{P}\left(|T^{(1)}_{2n+1} - \mu| \leq c^{(1)}_n\right) = 0.95\\ \Rightarrow &\mathbb{P}\left(\frac{|T^{(1)}_{2n+1} - \mu|}{\sigma\sqrt{2n+1}} \leq \frac{c^{(1)}_n}{\sigma\sqrt{2n+1}}\right) = 0.95\\ \Rightarrow &c^{(1)}_n = \sigma\sqrt{2n+1}\phi^{-1}(0.975) \approx 1.96\sigma\sqrt{2n+1}.\end{aligned}$$

For the second estimator, define $\tilde{X}_k := X_k - \mu \sim N(0,\sigma)$ and $\tilde{X}_{(k)} = (\tilde{X})_{(k)}$, then $F^{-1}\left(\frac{1}{2}\right) = 0$. Thanks to the example 4.4 of the Skript, 2e know that:

$$\mathbb{P}\left(\sqrt{2n+1}\tilde{X}_{(n+1)} \leq x\right) \to \phi(2F'(0)x),$$

where in our case $F'(0) = \frac{1}{\sqrt{2\pi}\sigma}$. Then,

$$\mathbb{P}\left(|T_n^{(2)} - \mu| \leq x\right) = \mathbb{P}\left(\sqrt{2n+1}\tilde{X}_{(n+1)} \leq \sqrt{2n+1}x)\right) + \mathbb{P}\left(\sqrt{2n+1}\tilde{X}_{(n+1)} \geq -\sqrt{2n+1}x\right)$$
$$\approx 1 - 2\phi\left(\frac{\sqrt{2}}{\sqrt{\pi}\sigma}\sqrt{2n+1}x\right),$$

then if we take $c_n^{(2)} := \phi^{-1}(97.5)\frac{\sqrt{\pi}}{\sqrt{2}\sqrt{2n+1}}\sigma$ we have what we wanted.

(b) Taking $q = \frac{\pi}{2}$ we have that:

$$\frac{c_{qn}^{(2)}}{c_n^1} \approx \frac{\sqrt{\pi}}{\sqrt{2}}\frac{\sqrt{2n+1}}{\sqrt{\pi n+1}} \to 1.$$

The parameter q represents how many more data I have to take with the estimator 2 to get the same order of error bounds than for the one of experiment 1.

Q4. A gas station estimates that it takes at least α minutes for a change of oil. The actual time varies from costumer to costumer. However, one can assume that this time will be well represented by an exponential random variable. The random variable X, therefore, possess the following density funciont

$$f(t) = e^{\alpha - t}\mathbf{1}_{\{t \geq \alpha\}},$$

i.e. $X = \alpha + Z$ where $Z \sim Exp(1)$. The following values were recorded from 10 clients randomly selected (the time is in minutes):

$$4.2, 3.1, 3.6, 4.5, 5.1, 7.6, 4.4, 3.5, 3.8, 4.3.$$

Estimate the parameter α using the estimator of maximum likelihood.

Solution:

We have that the likelihood function is given by:

$$L(X_1, ..., X_n, \alpha) = \prod_{i=1}^{n} \exp(\alpha - X_i)\mathbf{1}_{\{X_i \geq \alpha\}},$$
$$= \exp(n\alpha - \sum_{i=1}^{n} X_i)\mathbf{1}_{\{\bigcap_{i=1}^{n} X_i \geq \alpha\}},$$

we note that $f(\alpha) := \exp(n\alpha - \sum_{i=1}^{n} X_i) > 0$ is increasing, so its maximum is attained at the maximum point where $\mathbf{1}_{\{\bigcap_{i=1}^{n} X_i \geq \alpha\}} \neq 0$. Then the point that maximizes the likelihood is in $\bar{\alpha} = \min_{i=1,..,n}\{X_i\}$.