EPFL, Spring 2023
Guillaume Obozinski

EE-209 Eléments de Statistiques pour les Data Sciences

Feuille d'exercices 10

Exercise 10.1 Prove that the *sample empirical covariance* s_{xy} satisfies the following identities

$$(n-1)s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n} x_i(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}.$$

Exercise 10.2 An industrial machine is used repeatedly to cut small metal sheets of desired length. Unfortunately, as the machine is being used, because of drifts in position between mechanical parts of the machine, the precision of the machine degrades and the length of the sheets produced can be off by a certain amount.

Hours of machine use	150	165	170	175	195	220	225
Discrepancy (in mm)	1.10	1.21	1.25	1.23	1.30	1.40	1.42

(a) Write the formulas to compute $\widehat{a}$ and $\widehat{b}$ from the data $(t_1, y_1), \dots, (t_7, y_7)$, and show that

$$\widehat{a} = \frac{\sum_i t_i y_i - n\bar{t}\,\overline{y}}{\sum_i t_i^2 - n\bar{t}^2}.$$

Why can this formula be convenient to perform calculations?

(b) Determine the linear equation $y = \widehat{a}t + \widehat{b}$ relating discrepancy y to time t using linear regression.

Exercise 10.3 We consider a situation in which we would like to model the relationship between an explanatory variable (e.g., the distance from the sea) and a response variable (e.g., the pluviometry). The sites where the measurements of rain will be made have been decided by experts, given a number of practical constraints, and so their distances to the sea $x_1, \dots, x_n$ are not considered random, but simply fixed values. However, the amount of rain measured at each site will effectively be well modelled by a random variable. Previous work has established that a linear model is reasonable so we model the response as a random variable $Y_i = ax_i + b + \varepsilon_i$, where a, b and x_i are fixed quantities, and ε_i are i.i.d. random variables, with zero expectation and with the same variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.

(a) What are $\mathbb{E}[Y_i]$ and $\mathrm{Var}(Y_i)$ equal to?

(b) What is $\mathbb{E}[\widehat{a}]$ equal to? What does this mean for the estimator $\widehat{a}$?

(c) Calculate $\mathrm{Var}(\widehat{a})$. Hint: Express $\widehat{a}$ as a function of $a, x_i, \varepsilon_i, \overline{x}$ and use a well chosen formula from the first problem, to write it as a sum of independent r.v.s.

(d) We assume in this question that the ε_is are i.i.d. centered (i.e. such that $\mathbb{E}[\varepsilon_i] = 0$) *Gaussian* random variables with known variance σ^2. Determine the form of a 95% symmetric confidence interval for a based on $\widehat{a}$, σ, n, and s_x^2.

(e) Obviously, in most cases the variance σ^2 is unknown. It can be shown that, if $\widehat{Y}_i = \widehat{a}x_i + \widehat{b}$ and if $e_i = Y_i - \widehat{Y}_i$ is the ith residual, then

$$\widehat{\sigma}_e^2 := \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

is an unbiased estimator of σ^2. The $n-2$ in the denominator might seem a bit surprising but this is because the two parameters $\widehat{a}$ and $\widehat{b}$ that are estimated remove two degrees of freedom. (Please note as well that ε_i is not the same thing as e_i: the first one is a random variable that we don't have access to while e_i are the calculated *residuals*). Furthermore it can be shown, under the same assumptions as in the previous question, namely that ε_is are i.i.d. centered (i.e. such that $\mathbb{E}[\varepsilon_i] = 0$) *Gaussian* random variables, with unknown variance σ^2, that

$$T := \frac{\sqrt{n-1}\, s_x}{\widehat{\sigma}_e}(\widehat{a} - a)$$

follows a Student distribution with $n-2$ degrees of freedom. Determine a Student confidence interval for a.

(f) Using the same ideas as in the previous question, and with the same assumptions on the distribution of ε, propose a Student test to test the null hypothesis $H_0 : a = 0$ vs $H_1 : a \neq 0$ at 5% significance level. In particular, specify the test statistic and the critical region.

Exercise 10.4 Assume that X and Y are two independent random variables with finite variance and with $\mathrm{Var}(X) > 0$. Based on a sample $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size n of i.i.d. observations following the same distribution as (X, Y), we denote by $\widehat{a}_n$ and $\widehat{b}_n$ the slope and intercept of the simple linear regression obtained by regressing y on x.

(a) What do you expect the values of $\widehat{a}_n, \widehat{b}_n$ and of the coefficient of determination r^2 to be close to ? Is it possible to formally say that they converge to some values? If so make the formal statement and prove it.

(b) Is there a weaker assumption than independence of X and Y under which the same conclusion are obtained.