Université Côte d'Azur

MSc 1 Data Science & IA
2022-2023

Statistical inference practice, part 2: mid-term exam

Duration 1h
Any kind of paper document allowed
Any kind of calculator allowed

Exercise 1 (5.5 points)

Let consider a Gaussian vector $\mathbf{y} = (y_1, y_2, y_3)'$ a random vector of $\mathbb{R}^3$ with vector of mean $\boldsymbol{\mu} = (1, 2, -1)$ and with covariance matrix $\Sigma = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$

1. Give cov(y_1, y_2), var(y_1)
2. Are y_1 and y_3 independent?
3. Give the value of linear correlation coefficient between y_1 and y_2
4. Give the distribution of $y_1 - 2y_2 + y_3$
5. Let define by $\mathbf{z} = (y_1 - 2y_2 + y_3, y_1 + y_2)'$ a bivariate random variable. Give the distribution of $\mathbf{z}$.
6. Give the expression of the density function of $\mathbf{y}$

Exercise 2 (3 points)

Let consider x_1, x_2, x_3, x_4 and x_5 five independent standard normal variables.
Give the distribution of:

1. $x_1^2 + x_2^2 + x_3^2$
2. $\frac{x_1}{\sqrt{\frac{x_2^2 + x_3^2 + x_4^2}{3}}}$
3. $\frac{\frac{x_1^2 + x_2^2}{2}}{\frac{x_3^2 + x_4^2 + x_5^2}{3}}$

Exercise 3 (2 points)

Suppose that $y_1, y_2, \ldots, y_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$ let

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- What is the distribution of $(n-1)\frac{s^2}{\sigma^2}$?
- What is the distribution of $t = (\bar{y} - \mu)/(s/\sqrt{n})$?

Exercise 4 (5.5 points)

Let assume the following model

$$y_i = \beta x_i + \epsilon_i, \ i = 1, 2, \ldots, n$$

with :

- $E(\epsilon_i) = 0$ for all $i = 1, 2, \ldots, n$
- $\text{var}(\epsilon_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$
- $\text{cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$

1. What is the link between this model and the simple linear regression model?
2. Show that the least-square estimator of β is:
$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$
3. Compute $E(\hat{\beta})$ and deduce that $\hat{\beta}$ is an unbiased estimator of β
4. Compute the variance of $\hat{\beta}$
5. Let now add the assumption that the ϵ_i $(i = 1, 2, \ldots, n)$ follow a Gaussian distribution, deduce the distribution of $\hat{\beta}$
6. Propose and estimator of σ^2

Exercise 5 (4 points)

Assume that we have fitted a multiple linear regression on $n = 100$ data point with y as response and $x1$, $x2$, $x3$ and $x4$ as covariates, here is the summary of the model fit :

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44933 -0.10983 -0.01174  0.10368  0.30816

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.044052   0.016015   2.751  0.00712 **
```

```
x1           2.992544   0.020010 149.556  < 2e-16 ***
x2           1.948712   0.017065 114.195  < 2e-16 ***
x3          -1.035952   0.018936 -54.709  < 2e-16 ***
x4           0.004175   0.019105   0.219  0.82747
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1597 on 95 degrees of freedom
Multiple R-squared:  0.998,Adjusted R-squared:  0.9979
F-statistic: 1.168e+04 on 4 and 95 DF,  p-value: < 2.2e-16
```

1. Give the value of $\hat{\beta}_1$
2. Give the value of $\widehat{\sigma(\beta_1)}$
3. What is the value of R^2 ?
4. The residual standard error is given by the formula:

$$\sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

   with here $n - k - 1 = 95$. Thus deduce the estimed value of σ^2 (the variance of the noise in the linear model).