Time: 1 hour 10 mins

In the TESTO folder, you can find the RMarkdown file "consegna.Rmd". Write your answers (R code and text) there, then

1. use the Knit button to generate an HTML file
2. name the HTML file with your badge number
3. upload the HTML file to the CONSEGNA folder

Other formats will not be accepted.

Exercize 1

Points 3

Longley's Economic Regression Data is a macroeconomic data set with 7 economical variables, observed yearly from 1947 to 1962 ($n = 16$). Type

```
longley # import data
?longley # help
```

to import the dataset and to get further information about the variables. The response variable is `GNP.deflator` and the 6 predictors are `GNP`, `Unemployed`, `Armed.Forces` , `Population`, `Year` and `Employed`.

a. **Print in output** the rigde estimate

$$\hat{\beta}(\lambda) = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda \mathbf{I}_{p\times p})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

for $\lambda = 0.005$, where $\mathbf{X}$ is the design matrix with the first column of 1, thus the intercept term is penalized as well.

```
# write here the R code
```

b. **Print in output** the diagonal elements of

$$\mathbb{V}\text{ar}(\hat{\beta}(\lambda)) = \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{W}_\lambda^\mathsf{T}$$

for $\lambda = 0.005$ and $\sigma^2 = 1$, where $\mathbf{W}_\lambda = [\mathbf{I}_{p\times p} + \lambda(\mathbf{Z}^\mathsf{T}\mathbf{X})^{-1}]^{-1}$-

```
# write here the R code
```

c. **Print in output** the LOOCV error for
   - the linear model including all 6 predictors
   - the ridge regression model with $\lambda = 0.005$.

```
# write here the R code
```

Comment the results.

Write here your answer.

Exercize 2

Points 3

Consider a Fixed-X setting where the response is generated according to the model

$$y_i = f(x_i) + \varepsilon_i$$

where

- $x_i = i, \quad i = 1, \ldots, n$
- $n = 10$
- the true regression function is $f(x_i) = 1$
- $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ with $\sigma = 10$.

Consider a polynomial regression model of degree d. **Print in output** the test prediction error $\text{ErrF} = \mathbb{E}(\text{MSE}_{\text{Te}})$ for $d = 0, 1, \ldots, 9$.

```
# write here the R code
```

Exercize 3 (ISLR, Chapter 6, Applied Exercize 10)

Points 4

Generate a simulated data set as follows:

```
set.seed(123)
n = 1000
p = 20
X = matrix(rnorm(p*n), ncol=p)
beta = c(2,rep(1,5),rep(0,15))
y = beta[1] + X%*%beta[-1] + rnorm(n)
```

where the true coefficients are $\beta_0 = 2$ (intercept), $\beta_1 = \ldots = \beta_5 = 1$ and $\beta_6 = \ldots = \beta_{20} = 0$.

Split your data set into a training set containing the first 100 observations and a test set containing the last 900 observations.

Perform best subset selection on the training set to obtain the best model of size k (i.e. the model including the intercept term and k selected predictors) for $k = 1, 2, \ldots, 10$.

Note that if the jth predictor is not included in the best model of size k, then $\hat{\beta}_j^k = 0$, where $\hat{\beta}_j^k$ is the jth coefficient estimate for the best model of size k, $j = 0, 1, \ldots, 20$.

Print in output the scatterplot displaying $d_k = \sqrt{\sum_{j=0}^{20} (\beta_j - \hat{\beta}_j^k)^2}$ (y-axis) for a range of values of k from 1 to 10 (x-axis).

```
# write here the R code
```

Comment on what you observe.

Write here your answer.