Data Mining

Name:

Surname:

Badge number:

Exercize I (ISLR, Chapter 5, Applied Exercize 8)

We will now perform cross-validation on a simulated data set.

a. Generate a simulated data set as follows:

```r
set.seed(1)
x = rnorm(100)
y = x - 2*x^2 + rnorm(100)
```

In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

Write here your answers.

b. Create a scatterplot of X against Y. Comment on what you find.

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

Write here your comments.

c. Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:

(i). $Y = \beta_0 + \beta_1 X + \epsilon$

(ii). $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

(iii). $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

(iv). $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

Note you may find it helpful to use the data.frame() function to create a single data set containing both X and Y.

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

d. Repeat c. using another random seed, and report your results. Are your results the same as what you got in c.? Why?

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

Write here your answers.

e. Which of the models in c. had the smallest LOOCV error? Is this what you expected? Explain your answer.

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

Write here your answers.

Exercize II

Consider a fixed-design setting with $n = 21$,

$$x_i = -2 + (i-1)0.2, \quad i = 1, \dots, n$$

and true regression function a polynomial of degree 5

$$f(x_i) = \frac{1}{20}(x_i+4)(x_i+2)(x_i+1)(x_i-1)(x_i-3)+2$$

Then

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

with $\epsilon_i \sim N(0, \sigma^2)$ with $\sigma = 1$.

Suppose you are considering to use a polynomial regression model of degree $d = 1, 2, \ldots, 10$ and you want to select the best degree d^* which minimizes the prediction error $\mathrm{ErrF} = \mathbb{E}(\mathrm{MSE}_{\mathrm{Te}})$.

a. Plot $(x_i, f(x_i))$ for $i = 1, \ldots, n$.

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

b. Print in output the squared bias for each degree d

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

c. Print in output the variance for each degree d

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

d. Which is the degree d^* that minimize ErrF? Is this what you expected? Explain your answer.

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

Write here your answer.