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:

```
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.

```
# 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.

```
# 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?

```
# 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.

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

Write here your answers.