Lab 6 – Practice Questions

Q1)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)
```
b) Create a scatterplot of X against Y . Comment on what you find.
c) Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:
i. $Y = \beta_0 + \beta_1 X + \varepsilon$

ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon$.
d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?
e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.
f) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?