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

(a) Generate a simulated data set as follows.

```
set.seed(1)
y=rnorm (100)
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.

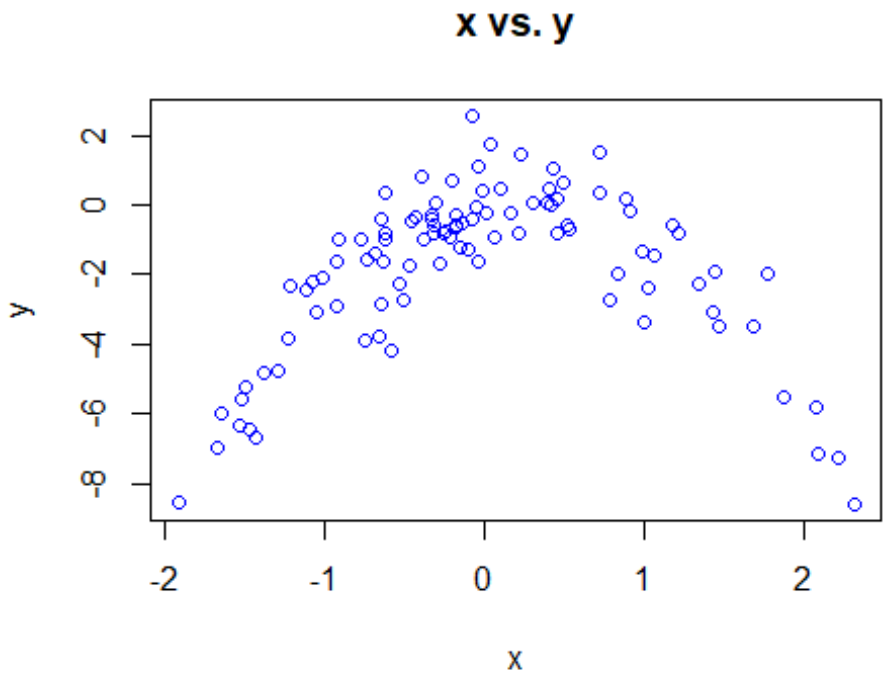
Tenemos que $n = 100$ y $p = 2$, el modelo usado para generar esta data es:

$$y = x - 2x^2 + \varepsilon$$

$$\varepsilon \sim N(0,1)$$

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

```
plot(x,y, main="x vs. y", col="blue")
```



Se observa claramente una relación curva **parabólica** entre **x** y **y**

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

```
library(boot)
### se crea un data frame para facilitar la manipulación de los datos
Data <- as.data.frame(cbind(x, y))
set.seed(1)
```

$$i. Y = \beta_0 + \beta_1 X + \varepsilon$$

```
glm.fit1 <- glm(y ~ x)
cv.error1<-cv.error<-cv.glm(Data, glm.fit1)$delta
cv.error

## [1] 5.890979 5.888812
```

La función cv.error<-cv.glm() produce una lista con varios componentes. Los dos números del vector delta contienen los resultados de la validación cruzada. En este caso los números son idénticos (hasta dos decimales) y corresponden a la estadística LOOCV.

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

```
glm.fit2 <- glm(y ~ poly(x, 2))
cv.error2<-cv.glm(Data, glm.fit2)$delta
cv.error2

## [1] 1.086596 1.086326
```

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

```
glm.fit3 <- glm(y ~ poly(x, 3))
cv.error3<-cv.glm(Data, glm.fit3)$delta
cv.error3

## [1] 1.102585 1.102227
```

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

```
glm.fit4 <- glm(y ~ poly(x, 4))
cv.error4<-cv.glm(Data, glm.fit4)$delta
cv.error4

## [1] 1.114772 1.114334
```

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

```
set.seed(10)
glm.fit1 <- glm(y ~ x)
cv.error1<-cv.glm(Data, glm.fit1)$delta
cv.error1

## [1] 5.890979 5.888812
```

```
glm.fit2 <- glm(y ~ poly(x, 2))
cv.error2<-cv.glm(Data, glm.fit2)$delta
cv.error2

## [1] 1.086596 1.086326

glm.fit3 <- glm(y ~ poly(x, 3))
cv.error3<-cv.glm(Data, glm.fit3)$delta
cv.error3

## [1] 1.102585 1.102227

glm.fit4 <- glm(y ~ poly(x, 4))
cv.error4<-cv.glm(Data, glm.fit4)$delta
cv.error4

## [1] 1.114772 1.114334
```

Los resultados son idénticos debido a que LOOCV solo usa una observación para validar el modelo, las demás observaciones son usadas para el entrenamiento.

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

El error cuadrático medio **MSE** más pequeño corresponde estimado por el LOOCV el modelo **glm.fit2**, puesto que como se establecido al durante la simulación del ejercicio la relación entre las variable es cuadrática.

(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?

```
summary(glm.fit4)

##
## Call:
## glm(formula = y ~ poly(x, 4))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8914  -0.5244   0.0749   0.5932   2.7796
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.8277     0.1041 -17.549   <2e-16 ***
## poly(x, 4)1   2.3164     1.0415   2.224   0.0285 *
## poly(x, 4)2 -21.0586     1.0415 -20.220   <2e-16 ***
## poly(x, 4)3  -0.3048     1.0415  -0.293   0.7704
```

```
## poly(x, 4)4  -0.4926     1.0415  -0.473   0.6373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.084654)
##
##     Null deviance: 552.21  on 99  degrees of freedom
## Residual deviance: 103.04  on 95  degrees of freedom
## AIC: 298.78
##
## Number of Fisher Scoring iterations: 2
```

Gracias al p-value se identifica que solo los coeficientes β_1 y β_2, que acompañan a los términos linear y cuadratico respectivamente son significativos. Tanto el **AIC** como la **deviación residual muestran** un buen ajuste del modelo, dado que sus valores no son muy grande.