4.

a.

It is clear that if $x \in [0.05, 0.95]$ then the observations we will use are in the interval $[x-0.05, x+0.05]$ and consequently represents a length of 0.1 which represents a fraction of 10%. If $x < 0.05$, then we will use observations in the interval $[0, x+0.05]$ which represents a fraction of $(100x+5)\%$; by a similar argument we conclude that if $x > 0.95$, then the fraction of observations we will use is $(105-100x)\%$. To compute the average fraction we will use to make the prediction we have to evaluate the following expression

$$\int_{0.05}^{0.95} 10 dx + \int_{0}^{0.05} (100x+5) dx + \int_{0.95}^{1} (105-100x) dx = 9 + 0.375 + 0.375 = 9.75.$$

So we may conclude that, on average, the fraction of available observations we will use to make the prediction is 9.75%.

b.

If we assume X_1 and X_2 to be independant, the fraction of available observations we will use to make the prediction is $9.75\% \times 9.75\% = 0.950625\%$.

c.

With the same argument than (a) and (b), we may conclude that the fraction of available observations we will use to make the prediction is $9.75\%^{100} \simeq 0\%$.

d.

As we saw in (a)-(c), the fraction of available observations we will use to make the prediction is $(9.75\%)^p$ with p the number of features. So when $p \to \infty$, we have

$$\lim_{p \to \infty} (9.75\%)^p = 0.$$

e.

For $p = 1$, we have $l = 0.1$, for $p = 2$, we have $l = 0.1^{1/2}$ and for $p = 100$, we have $l = 0.1^{1/100}$.

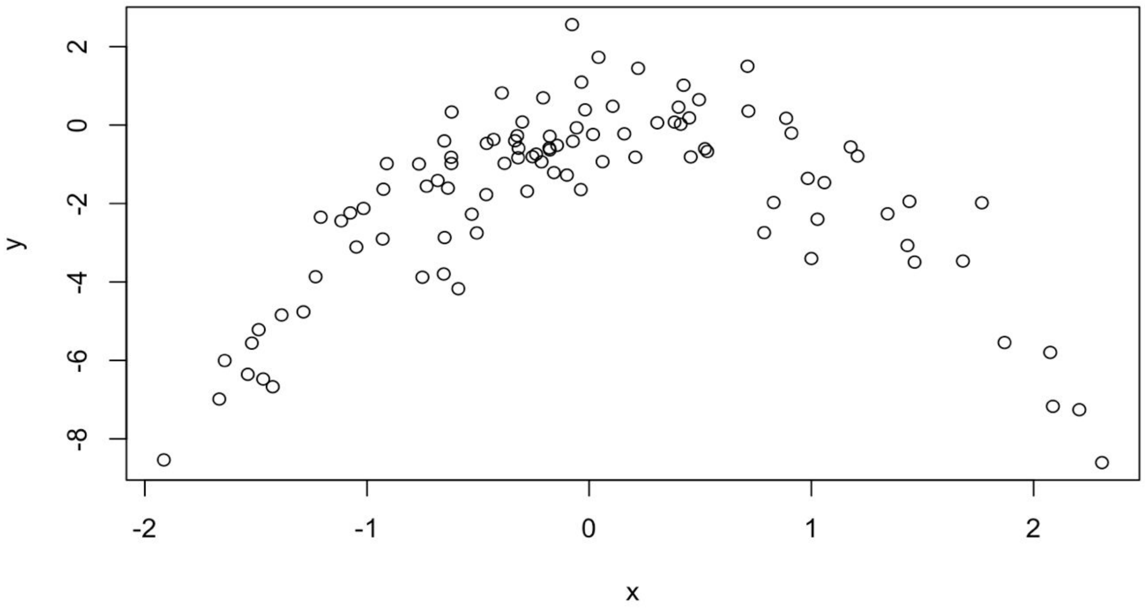
6.

(a)

Here we have that $n = 100$ and $p = 2$, the model used is

$$Y=X-2*x^2+\varepsilon$$

(b)



The data obviously suggests a curved relationship.

(c) & (d)

The results above are identical to the results obtained in (c) since LOOCV evaluates nn folds of a single observation.

(e)

We may see that the LOOCV estimate for the test MSE is minimum for “fit.glm.2”, this is not surprising since we saw clearly in (b) that the relation between “x” and “y” is quadratic.

The p-values show that the linear and quadratic terms are statistically significant and that the cubic and 4th degree terms are not statistically significant. This agree strongly with our cross-validation results which were minimum for the quadratic model.