M339G: November 11th, 2024.

0.632

Say, we are doing bootstrap.

Let our original sample be $x_1, x_2, \dots, x_n$

With bootstrap, we draw with replacement from the original sample.

Focusing on, say x_1, the probability of it not being chosen in one draw is

$$\underline{1 - \frac{1}{n}}$$

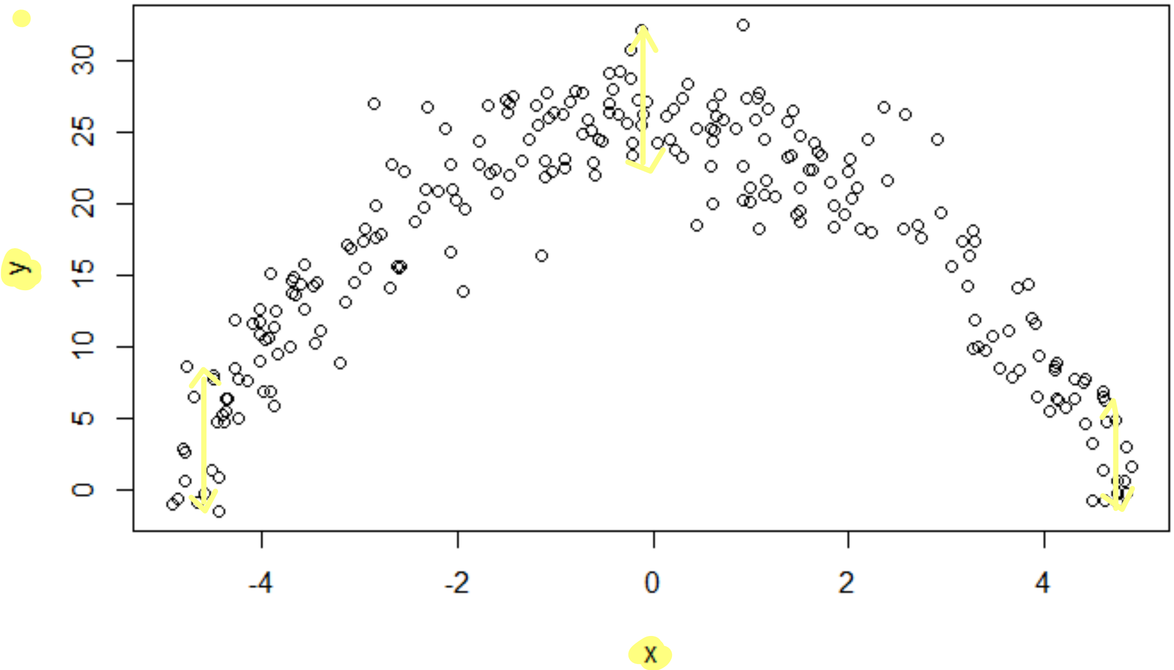
But, there are n independent draws. So, the total probability of never choosing x_1 is

$$\underline{\left(1 - \frac{1}{n}\right)^n} \xrightarrow[n \to \infty]{} \frac{1}{e} = e^{-1} \approx 0.368$$

So, $1 - e^{-1} = 0.632$ is the proportion (on average) of the data points that do end up in the bootstrapped sample.

39. You are given a dataset with two variables, which is graphed below. You want to predict *y* using *x*.

Determine which statement regarding using a generalized linear model (GLM) or a random forest is true.



X (A) A random forest is appropriate because the dataset contains only quantitative variables. Trees in general work well w/ qualitatie predictors.

X (B) A random forest is appropriate because the data does not follow a straight line. The opposite is true.

X (C) A GLM is not appropriate because the variance of *y* given *x* is not constant. The variance looks pretty constant.

X (D) A random forest is appropriate because there is a clear relationship between *y* and *x*. Could be anything!

(E) A GLM is appropriate because it can accommodate polynomial relationships.