Cross Validated

Stamatics, IIT Kanpur

Alok Kumar (190099), Tarun Kanodia (190902), Nilay Beniwal (190555)

Mini-Project 1

Problem Statement

For $i = 1, ..., n$, let $x_i = (1, x_{i2}, ..., x_{ip})$ be the vector of covariates for the ith observation and $\beta \in R_p$ be the corresponding vector of regression coefficients. Suppose response y_i is a realization of Y_i with

$$Y_i = \text{Bern}(\phi(x_i^T \beta))$$

where $\phi(.)$ is the CDF of a standard normal distribution.

1. Write an algorithm for obtaining the maximum likelihood estimator of β with all mathematical details.

2. On April $10, 1912$, the RMS Titanic sank after colliding with an iceberg on its maiden voyage. The accident killed 1502 of the 2224 passengers and crew on board.

   A subset of the data of the survival of passengers and crew can be loaded in R using titanic <- read.csv("titanic.csv").

   Note: You must have the dataset saved in the same folder as R/Python script to load the dataset.

   The dataset has the response: whether a passenger survived the tragedy (Survived). Additional information includes the sex of the passenger (1 if male, 0 otherwise)(Sexmale), age (Age), the number of siblings/spouse aboard (SibSp), the number of parents/children aboard (Parch) and the passengers fare (Fare) (in British pounds). The dataset contains 712 observations. Obtain the MLE estimates of the corresponding β coefficients using the model described above.

3. Jack Dawson was 20 years old when the tragedy happened. He boarded the ship with no family or spouse and paid 7.5 British pounds for his ticket. Rose Bukater was 19 years old on the tragic day. She boarded the ship with her fiance (treat this as a spouse) and her mother. She paid 512 British pounds for her ticket. What are the estimated probabilities of survival for Jack Dawson and Rose Bukater from your solution in part (b)?

Solution

The probability that any response takes the value 1 is:

$$\begin{aligned} p_i &:= \phi(x_i^T \beta) \\ &= \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(x_i^T \beta)^2} \end{aligned}$$

The likelihood function is given by:

$$L(\beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{(1-Y_i)}$$

It is easier to work with the log likelihood function. Hence:

$$\begin{aligned} \ell(\beta_1, \ldots, \beta_p) &= \sum_{i=1}^{n} [Y_i log(p_i) + (1 - Y_i) log(1 - p_i)] \\ &= \sum_{i=1}^{n} [Y_i \{log(\frac{1}{\sqrt{2\pi}}) - \frac{1}{2}.(x_i^T \beta)^2\} + (1 - Y_i)\{\log(1 - \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(x_i^T \beta)^2}\}] \end{aligned}$$

Taking the partial derivative with respect to β_m:

$$\frac{\partial \ell}{\partial \beta_m} = \sum_{i=1}^{n} Y_i \{(-x_i^T \beta)(x_{im}) + (1 - Y_i) \frac{1}{1 - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i^t \beta)^2}} . \frac{-1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i^T \beta)^2} . (-x_i^T \beta) . (x_{im})$$

$$= \sum_{i=1}^{n} (x_i^T \beta) x_{im} [-Y_i + (1 - Y_i).\frac{K}{1-K}]$$

where $K := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i^T \beta)^2}$

Hence, the derivative is:

$$\nabla \ell = \sum_{i=1}^{n} (x_i^T \beta) x_i [-Y_i + (1 - Y_i) \frac{K}{1-K}] = 0$$

We set the above equation to zero, so as to obtain **MLE** of β.An analytical solution is not possible here, thus a numerical optimization tool is required. Since the likelihood function is concave, we have used **Gradient Ascent** for obtaining maximum likelihood estimate of β.

The MLE of β obtained for this dataset is:
$[-0.23799673,\ 1.13993185,\ 0.65691815,\ 0.37675694,\ 0.14453475,\ -0.4124524]$.

Once β is obtained, we have used it calculate the probabilities of survival of Jack Dawson and Rose Bukater. The probabilities estimated are as follows:

Probability of survival of Jack is : 0.2280661248269023

Probability of survival of Rose is : 0.36865661591460586