Local Poisson Regression

Joel Cantero Priego & Ricard Meyerhofer Parra

14/12/2019

Introduction

In this assignment we are going to modify the already implemented functions *h.cv.sm.binomial* and *loglik.CV* to obtain a bandwidth choice method for the local Poisson based on loo-CV estimation of the expected likelihood of an independent observation. Finally, we are going to apply it to the Country Development dataset.

Bandwidth Choice for Local Possion Regression

The loo-CV of the expected log-likelihood of an independent observation can be written like this when using h as bandwidth.

$$l_{CV}(h) = \frac{1}{n} \sum_{i=1}^{n} log(\hat{Pr}_h^{(-i)}(Y = y_i | X = x_i))$$

where $\hat{Pr}_h^{(-i)}(Y = y_i | X = x_i)$ is an estimation of

$$Pr(Y = y_i | X = x_i) = e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!}$$

and should be estimated by maximum local likelihood

$$\lambda_i = \mathbb{E}(Y | X = x_i)$$

To implement this function, we had to modify the loo-CV estimation of the expected log-likelihood as aforementioned which gave us the following code:

```r
loglik.CV <- function(x,y,h){
    n <- length(x)
    pred <- sapply(1:n,
                  function(i,x,y,h){
                    sm.poisson(x=x[-i],y=y[-i],h=h,eval.points=x[i],display="none")$estimate
                  }, x,y,h)
    sum = 0
    for (i in 1:n){
      sum = sum + (log((exp(-pred[i]) * ((pred[i]**y[i])/factorial(y[i])))))/n
    }
    return(-1*sum)
}
```

Now, since we have the log-likelihood method for the Poisson distribution, we can compute the best bandwidth using it:

```r
h.cv.sm.poisson <- function(x,y,rg.h=NULL,l.h=10,method=loglik.CV){
  cv.h <- numeric(l.h)
  i <- 0
  gr.h <- exp( seq(log(rg.h[1]), log(rg.h[2]), l=l.h))
  for (h in gr.h){
    i <- i+1
    cv.h[i] <- method(x,y,h)
  }
  return(list(h = gr.h, cv.h = cv.h, h.cv = gr.h[which.min(cv.h)]))
}
```

Local Poisson regression for the Country Development Dataset

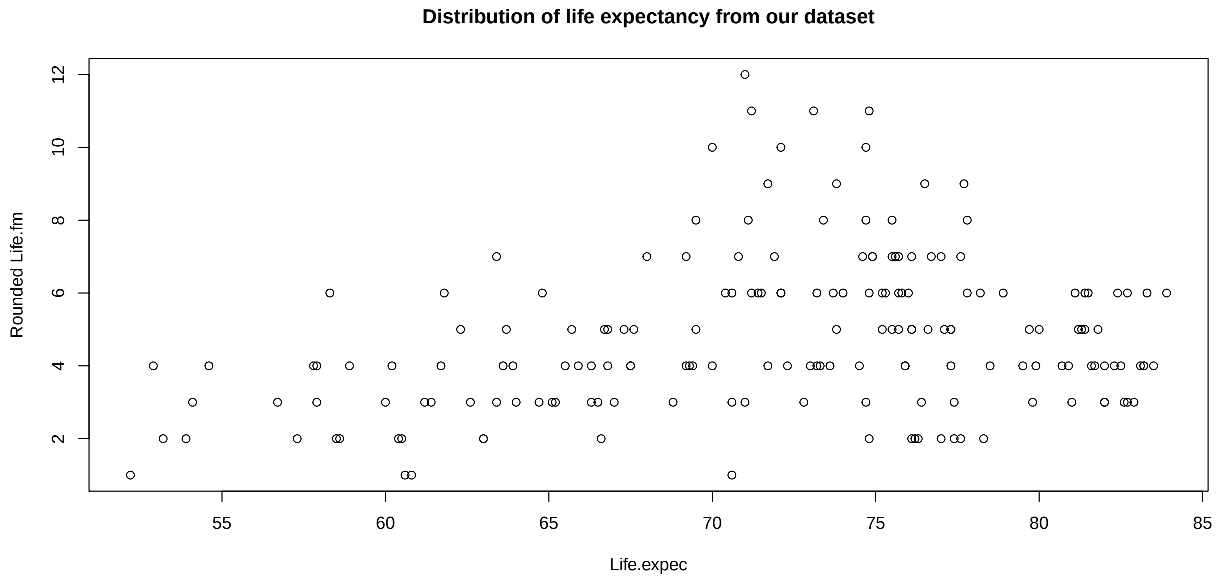
Now we are going to consider the country development dataset which contains information on development indicators for 179 countries. The parameters that we have are the following:

Variable name	Description	Values
iso3	Standard of country codes	String
country_name	Official Country name	String
Life.expec	Total life expectancy	Integer
Life.expec.f	Female life expectancy	Integer
Life.expec.m	Male life expectancy	Integer
le.fm	Result from Life.expec.f - Life.expec.m	Integer
Inf.Mort.rat	Mortality ratio	Integer
Agric.employ.%	Percentatge of employment that agriculture covers	Integer

As mentioned by the statement, we are going to load the dataset and we are going to create a new variable $le.fm.r$ which corresponds to the rounded value of $le.fm$ and our goal is to fit a local Poisson regression modeling $le.fm.0$ as a function of $Life.expec$.

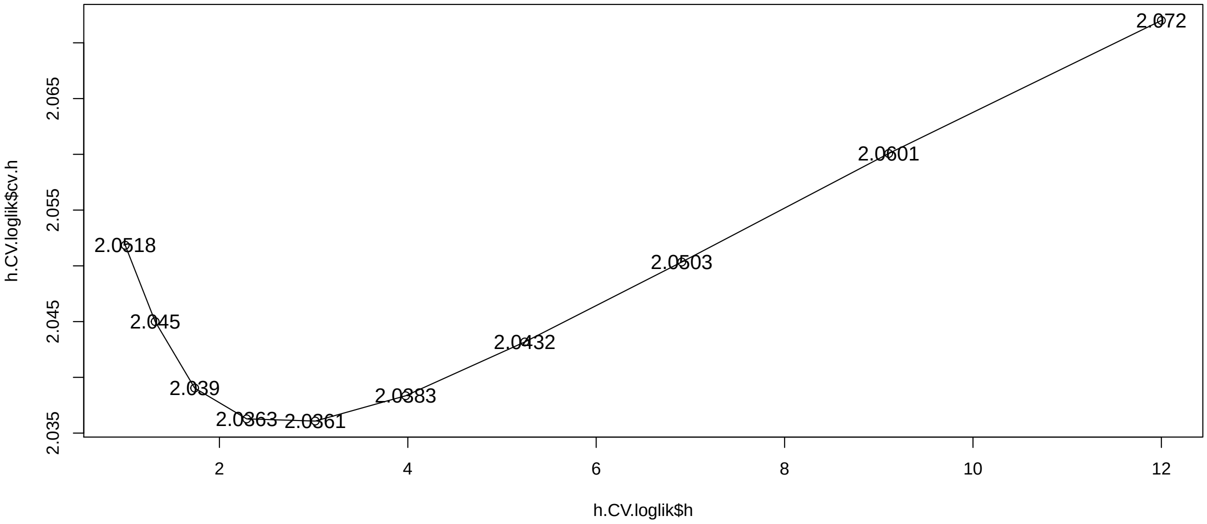
```
options(stringsAsFactors=F)
dataset <- read.csv("HDI.2017.subset.csv", sep = ";", header = T, dec=",")
dataset$le.fm.r <- round(dataset$le.fm)
```

```
plot(dataset$Life.expec, dataset$le.fm.r,main="Distribution of life expectancy from our dataset", xlab=
```

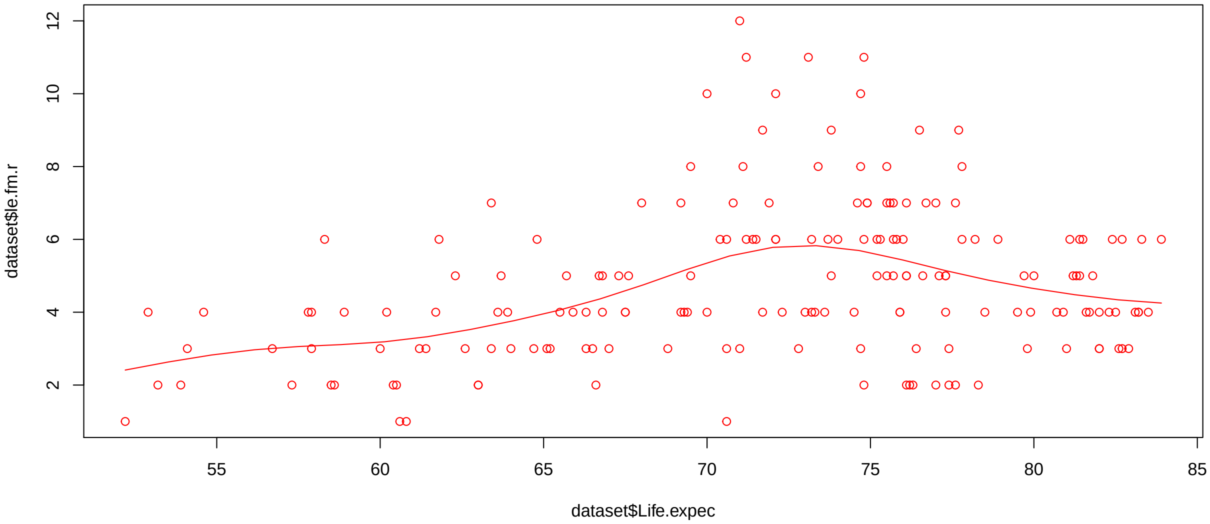


We are going to see which is the bandwidth that bests fits our model, we can see in the following graph that our minimum log-likelihood is found near 3 so is the value that we are going to use.

```
range.h <- c(1, 12)
h.CV.loglik <- h.cv.sm.poisson(dataset$Life.expec,dataset$le.fm.r,
                               rg.h=range.h,
                               method=loglik.CV)
```



In the below plot, we can see that the value we are using fits the data as we can see. The selected value is the one that was the closest from 3 which was *h.cv.loglik.h* = 3.017. It is not a perfect fit but the way the data is distributed does not seem it is possible to do better.


dataset$le.fm.r
dataset$Life.expec