

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 Poisson 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:

```
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:

```
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	Standar 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

```

options(stringsAsFactors=F)
dataset <- read.csv("HDI.2017.subset.csv", sep = ";", header = T, dec=",")
summary(dataset)

```

```

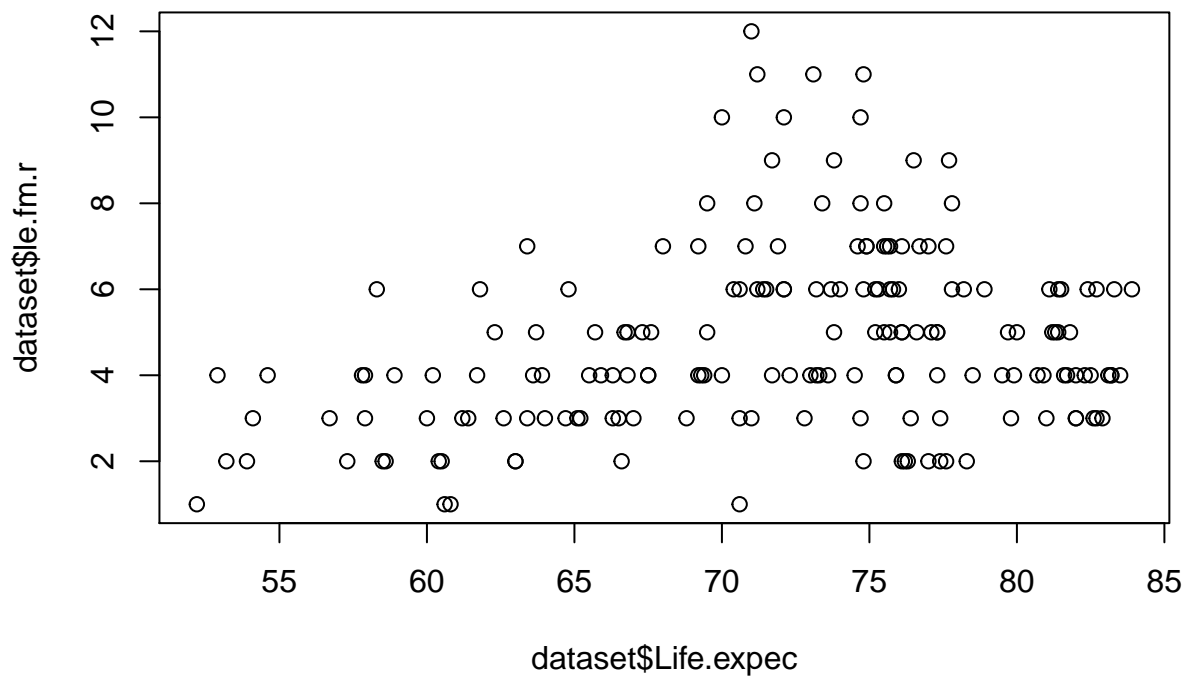
##      iso3          country_name      Life.expec      Life.expec.f
## Length:179      Length:179      Min.   :52.20      Min.   :52.80
## Class :character Class :character 1st Qu.:66.55      1st Qu.:68.25
## Mode  :character Mode  :character Median :73.40      Median :76.50
##                                     Mean  :71.85      Mean   :74.28
##                                     3rd Qu.:77.30      3rd Qu.:79.85
##                                     Max.   :83.90      Max.   :87.10
##      Life.expec.m      le.fm      Inf.Mort.rat      Agric.employ..
## Min.   :51.00      Min.   : 0.600      Min.   : 1.60      Min.   : 0.10
## 1st Qu.:64.55      1st Qu.: 3.350      1st Qu.: 6.60      1st Qu.: 6.25
## Median :70.40      Median : 4.500      Median :15.10      Median :19.00
## Mean   :69.45      Mean   : 4.821      Mean   :23.26      Mean   :27.57
## 3rd Qu.:74.90      3rd Qu.: 6.050      3rd Qu.:37.40      3rd Qu.:41.95
## Max.   :81.50      Max.   :12.000      Max.   :88.50      Max.   :91.50

```

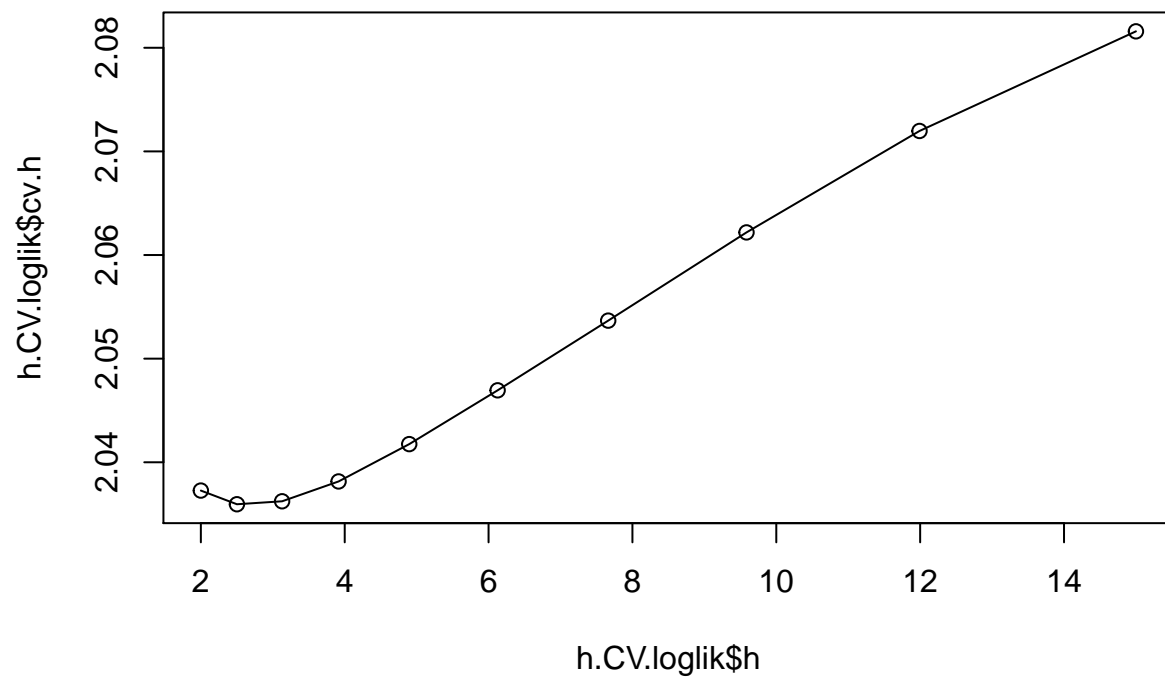
```

dataset$le.fm.r <- round(dataset$le.fm)
plot(dataset$Life.expec, dataset$le.fm.r)

```



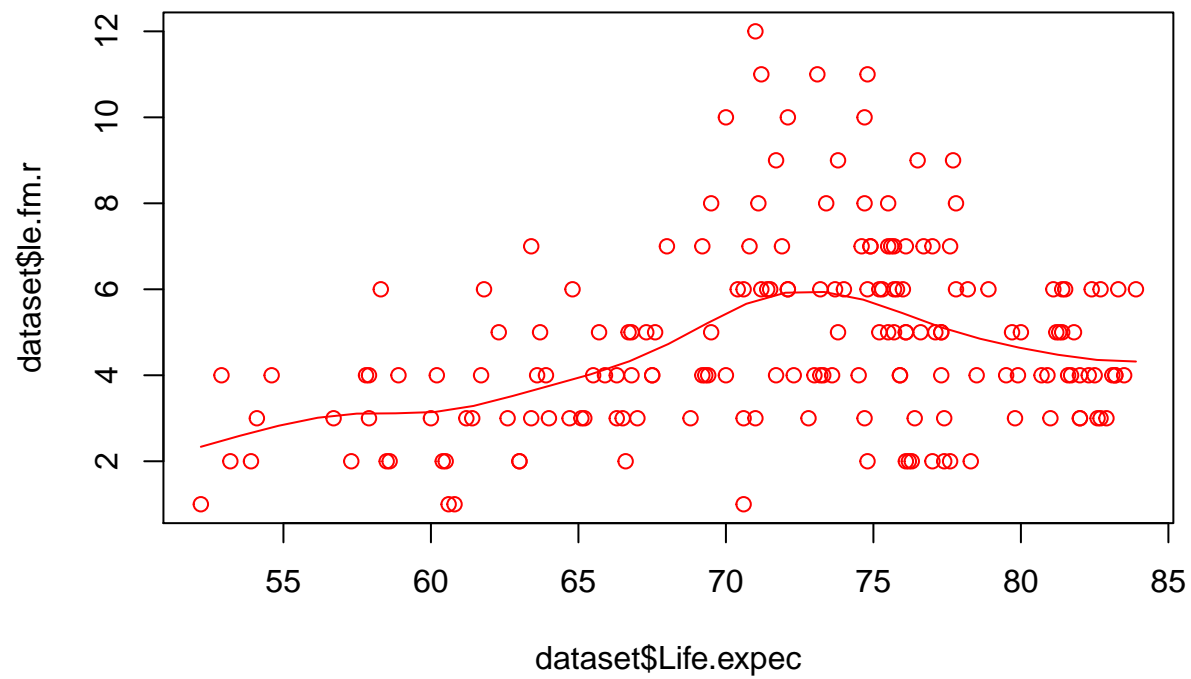
```
range.h = c(2, 15)
h.CV.loglik <- h.cv.sm.poisson(dataset$Life.expec,dataset$le.fm.r ,rg.h=range.h, method=loglik.CV)
plot(h.CV.loglik$h,h.CV.loglik$cv.h)
lines(h.CV.loglik$h,h.CV.loglik$cv.h)
```



```
kable(data.frame(h.CV.loglik$h, h.CV.loglik$cv.h))
```

h.CV.loglik.h	h.CV.loglik.cv.h
2.000000	2.037270
2.501837	2.035951
3.129594	2.036236
3.914868	2.038151
4.897181	2.041753
6.125974	2.046943
7.663094	2.053670
9.585907	2.062182
11.991188	2.071971
15.000000	2.081584

```
chosen.h = h.CV.loglik$h.cv
m1 <- sm.poisson(dataset$Life.expec, dataset$le.fm.r, h=chosen.h)
```



```
chosen.h
```

```
## [1] 2.501837
```