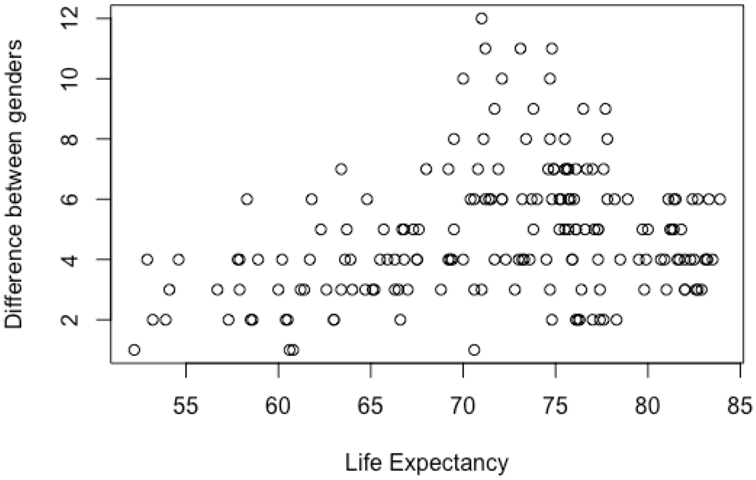
Local Poisson Regression

Marcel Pons Cloquells

December 21, 2020

1 Local Poisson Regression

In this homework, the Country Development dataset[1] will be used to analyze and obtain predictions on the population health status in 179 countries. More specifically, we are interested in getting insights about the differences on the life expectancy at birth between males and females having into account the overall life expectancy at birth for a given country.



As it can be appreciated in Figure 1, the relationship between the two variables does not follow a completely linear behavior in some values of x. For example, the differences in life expectancy between genders are higher at 70-75 but then lower in values greater than 75. By using a linear Poisson Regression (glm) this behaviour could not be appreciated and predictions would not be accurate for some cases of x.

To that end, a Local Poisson Regression using `sm.poisson()` will be performed using as explanatory variable `life.expec` and as response variable `le.fm.r`, which is the integer version obtained by rounding `le.fm` (differences in life expect. between genders).

Bandwidth choice

First of all, the best bandwidth value h has to be found. For this reason, the functions `h.cv.sm.poisson` and `loglik.CV` are built to obtain a bandwidth choice method for the local Poisson regression based on the loo-CV estimation of the expected likelihood of an independent observation.

[1] Source: Human Development Data (1990-2017), The Human Development Report Office, United Nations

The function `h.cv.sm.poisson` takes as input an explanatory variable x, a response variable y (integer), an optional range of bandwidth values $rg.h$[2] and the number of bandwidth values to be considered $l.h$.

The candidate list of bandwidth values $gr.h$ is obtained inside the function from $rg.h$ and $l.h$ and for every h in this list the `loglik.CV` function is called.

The `loglik.CV` performs a leave one out cross validation for every h using `sm.poisson` for $x^{(-i)}$ and $y^{(-i)}$ estimating the expected log likelihood for the independent observation (x_i). The mean of the expected log likelihoods for all the observations using every h is returned and stored in the list $cv.h$, which will be returned as one of the final outputs.

Formally, the leave one out cross validation for a given h is stated as:

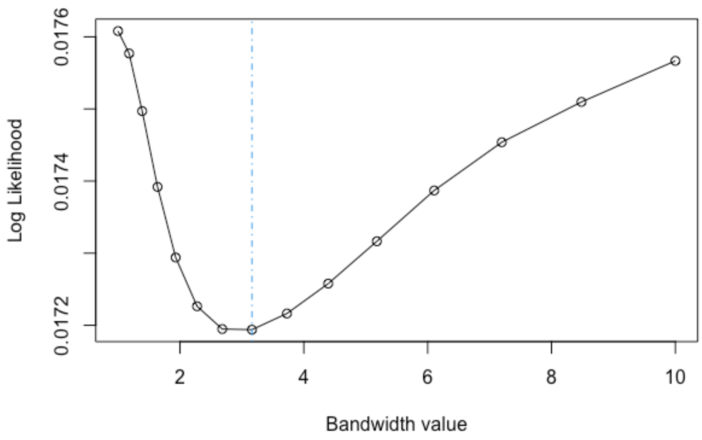
$$\ell_{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\widehat{\Pr}_h^{(-i)}(Y = y_i|X = x_i)\right),$$

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

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

Once the functions are defined[3], we execute them for `Life.expect` and `le.fm.r` considering a range from 1 to 10 and 15 bandwidth h values.

```
h.CV.loglik <- h.cv.sm.poisson(Life.expec,le.fm.r,rg.h=c(1,10),l.h=15)
```



It can be observed that the expected log likelihood is lower for the values of $h = 2.682$ and $h = 3.162$, being the later the value that gives the minimum log likelihood and the one that is going to be used for the Local Poisson Regression.

[2] If the range of bandwidth values is not specified in $rg.h$, the function `h.select()` is used to select the smoothing parameter for density estimation.

[3] The R code for the functions is provided in the Annex.

Applying the Local Poisson Regression

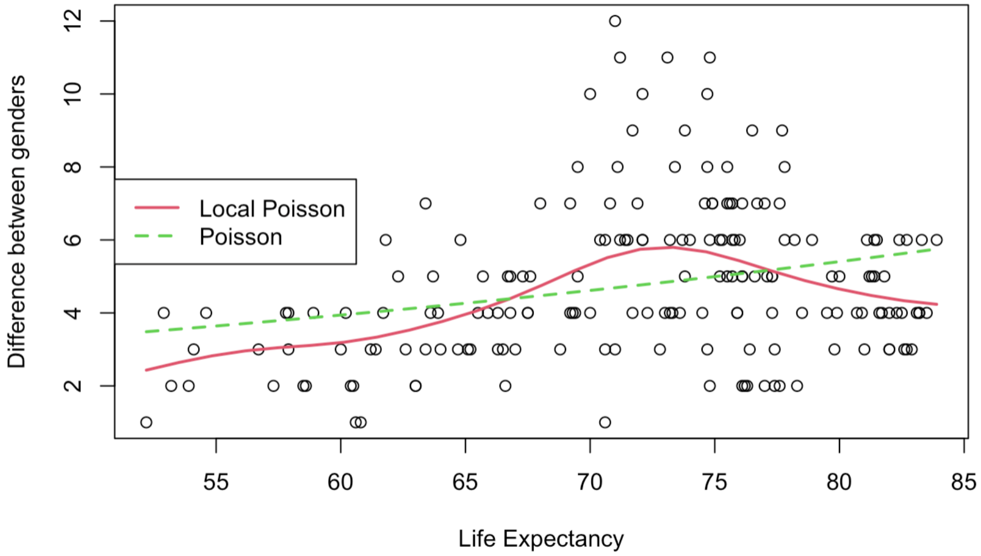
Now that we know which is the best bandwidth value h for the kernel function used for locally estimating `le.fm.r` given `Life.expec`, the Local Poission Regression is performed and compared with the Linear Poission Regression.

```
# Local Poisson Regression
aux <- sm.poisson(Life.expec, le.fm.r,h=h.CV.loglik$h.cv)

# Linear Poisson Regression
aux.glm <- glm(le.fm.r ~ Life.expec,family=poisson)
```

As it can be observed in Figure 3, the non parametric version of the Poisson Regression fits better the data than the linear Poisson Regression. By locally estimating `le.fm.r`, the non-linear behaviour of this variable is appreciated by the model, as it can be observed specially in the interval of life expectancy that goes from 70 to 85, where the differences between gender increase for 70-75 and then decrease for values greater than 75.

To sum up, the differences in life expectancy between genders are greater as the overall life expectancy increases, with the peak of highest differences at $\sim$ 74 and then for greater values the differences decreases.



2 Annex

Code of the developed functions

```r
h.cv.sm.poisson <- function(x,y,rg.h=NULL,l.h=10){
   cv.h <- numeric(l.h)
   if (is.null(rg.h)){
      hh <- c(h.select(x,y,method="cv"),
              h.select(x,y,method="aicc"))
      rg.h <- range(hh)*c(1/1.1, 1.5)
   }
   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] <- loglik.CV(x,y,h)
   }
   return(list(h = gr.h,
               cv.h = cv.h,
               h.cv = gr.h[which.min(cv.h)]))
}

loglik.CV <- function(x,y,h){
  n <- length(x)
  vec = numeric(n)
  for (i in range(1:n)){
     pred <-  sm.poisson(x[-i], y[-i],
                         h=h, display = "none",
                         eval.points = x[i])$estimate

     vec[i] <- -log(exp(-pred)*(pred^y[i])/factorial(y[i]))
  }
  return(sum(vec)/n)
}
```