ASM Practice

Local Poisson Regression

Maria Gkotsopoulou & Ricard Monge Calvo & Amalia Vradi

14/12/2019

The aim of this project is to obtain a function to choose the bandwith hyperparameter for local Poisson regression by using LOOCV estimates and apply it to the Country Development dataset [1].

This file contains the following variables:

- `Life.expec` Life expectancy at birth.
- `Life.expec.f` Life expectancy at birth for females.
- `Life.expec.m` Life expectancy at birth for males.
- `le.fm` Difference `Life.expec.f` minus `Life.expec.m`.
- `Inf.Mort.rat` Infant mortality rate: The annual number of deaths of infants under one year of age per 1,000 live births in the same year.
- `Agric.employ.%` Employment in agriculture (% of total employment).

We see the variable *le.fm* about life expectancy difference between male and female is always positive and can be rounded to give a Poisson-like distribution.

Bandwith choice functions for local Poisson regression

We modify the functions *h.cv.sm.binomial* and *loglik.CV* to obtain a bandwidth choice method for the local Poisson regression, based on the LOOCV estimation of the expected log-likelihood of an independent observation.

In the fuction *loglik.CV*, the process of fitting a *sm.poisson* model considering all datapoints apart from one is repeated n times. From each fitted model we obtain the estimates λ_i used to compute the log-likelihood. We use the following expression:

$$l_{CV} = \frac{1}{n}\sum_{i=1}^{n} \log\left(e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(-\lambda_i + y_i \log\lambda_i - \log y_i!\right)$$

[1] Human Development Data (1990-2017).

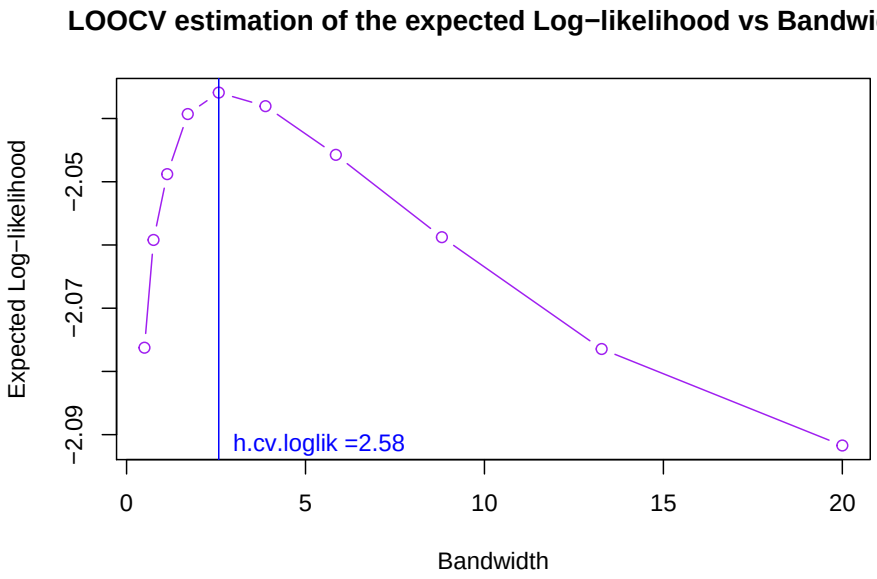
where n is the number of instances in the dataset, y_i is each of the response values and the expression $e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!}$ corresponds to the estimate of $Pr(Y = y_i|X = x_i)$, the probability of getting the response value given the predictors x_i and the estimate of the distributions parameter λ_i.

On the other hand, we rename the function *h.cv.sm.binomial* to *h.cv.sm.poisson* and change the selected bandwidth parameter *h.cv* as the one that maximizes the LOOCV estimates of the expected log-likelihood returned by the method function (*loglik.CV* in our case).

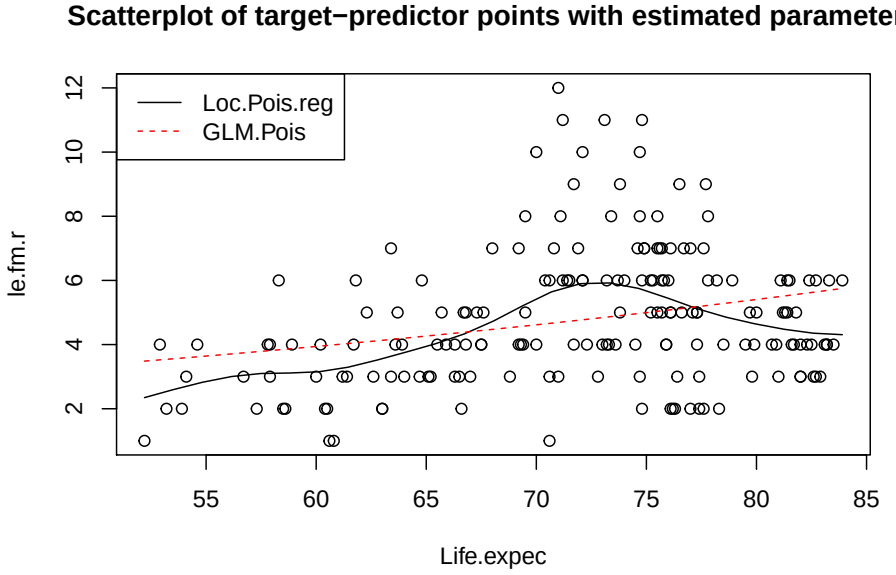
```
# Bandwidth choice in the local Poisson regression
# by leave-one-out cross-validation.
# Function "sm.poisson", from library "sm", is used.
# method can be equal to 'loglik.CV' or other method which needs to be MAXIMIZED
h.cv.sm.poisson <- function(x,y,rg.h=NULL,l.h=10,method=loglik.CV){
   cv.h <- 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] <- method(x,y,h)
   }
   return(list(h = gr.h,
               cv.h = cv.h,
               h.cv = gr.h[which.max(cv.h)])) # Maximizing log-likelihood
}
# method loglik.CV: leave-one-out log-likelihood
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)
  return(sum(y*log(pred) - pred - log(factorial(y)))/n)
}
```

Local Poisson regression for Country Development Data

We want to build a local Poisson regression to model life expectancy differnece between male and female *le.fm.r*, rounded to integer values, as a function of *Life.expec*, for the Country Development data. We use the previous function `h.cv.sm.poisson` to choose the bandwidth hyperparameter.



Assuming that the target variable, given the predictor, follows a Poisson distribution, our model estimates its mean rate λ. We show the scatterplot of the target-predictor points together with our model's estimated rates. In addition, we fit a Generalized Linear Model of the Poisson family, using the *glm* function, and plot its estimated rates for comparison purposes.



Estimates are similar between both models but the local Poisson regression seems to better fit to our sample.