Advanced Statistical Modeling

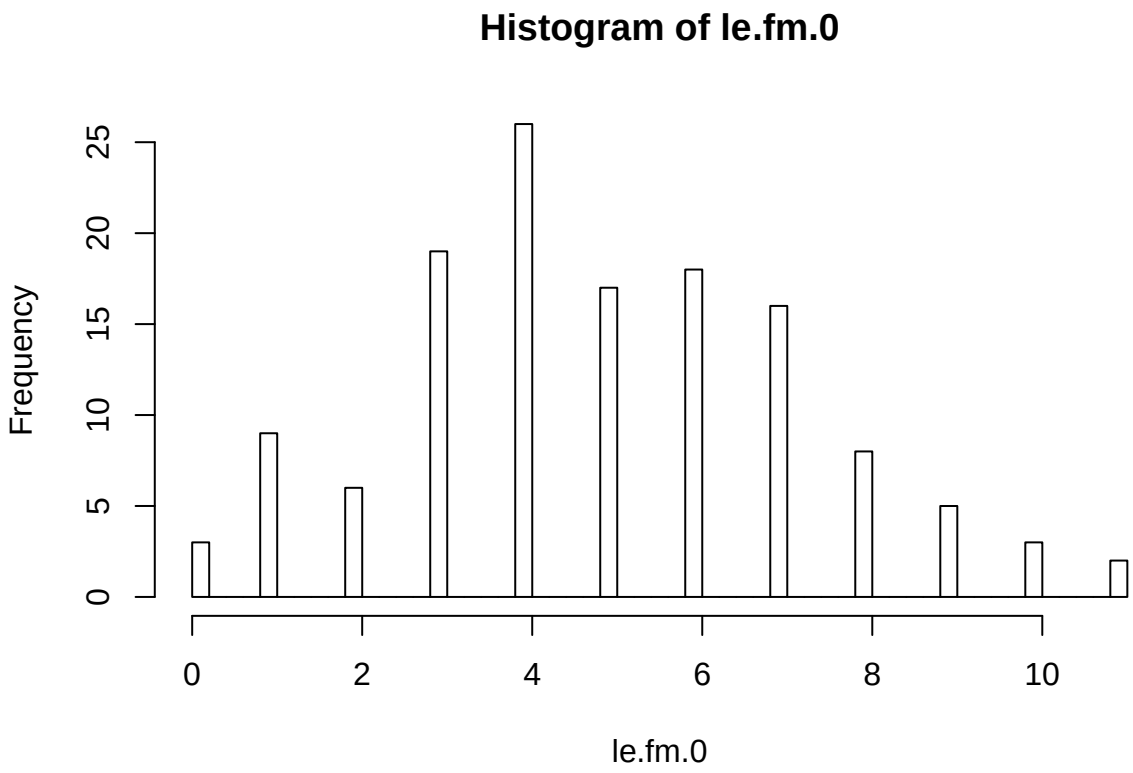
Non-parametric models - Generalized nonparametric regression model

Haoran Mo, Alexandra Yamaui

December 9th, 2017

In this exercise we are going to choose the bandwidth for a logistic regression model that has as response variable the difference in life expectancy between men and women (le.fm) in 132 countries. The data set used contains development indicators of the different countries. The variable le.fm always takes non-negative values, except for one country, so we are going to take 0 as minimum value.

```
countries<-read.table(file="countries.txt",head=T,row.names=2,dec=",")
attach(countries)
le.fm.0 <- pmax(0,le.fm)
hist(le.fm.0,br=40)
```



We can see that the frequency of difference of life expectancy between men and women shows a Poisson-like shape, it is bounded at 0, has discrete values and it is a little bit skewed to the left. Hence, will use local Poisson regression to estimate this variable

Fitting a local logistic regression involves choosing the right bandwidth, and in order to do that we are going to use the method of maximum log-likehood for new observations, which we will estimate using leave-one-out cross-validation. Below is the function of log-likelihood using h as bandwith

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

where $(y_i log(\frac{p_i^{(-i)}}{1-p_i^{(-i)}}) + log(1-p_i^{(-i)}))$ is the contribution to the log-likelihood function of each observation.

Additionally, $\hat{Pr}_h^{(-i)}(Y = y_i|X = x_i)$ is a estimation the probability of events for a Poisson distribution $Pr(Y = y_i|X = x_i) = e^{-\lambda_i}\frac{\lambda_i^{x_i}}{x_i!}$ and λ_i represents the average number of events per interval $\lambda_i = E(Y|X = x_i)$

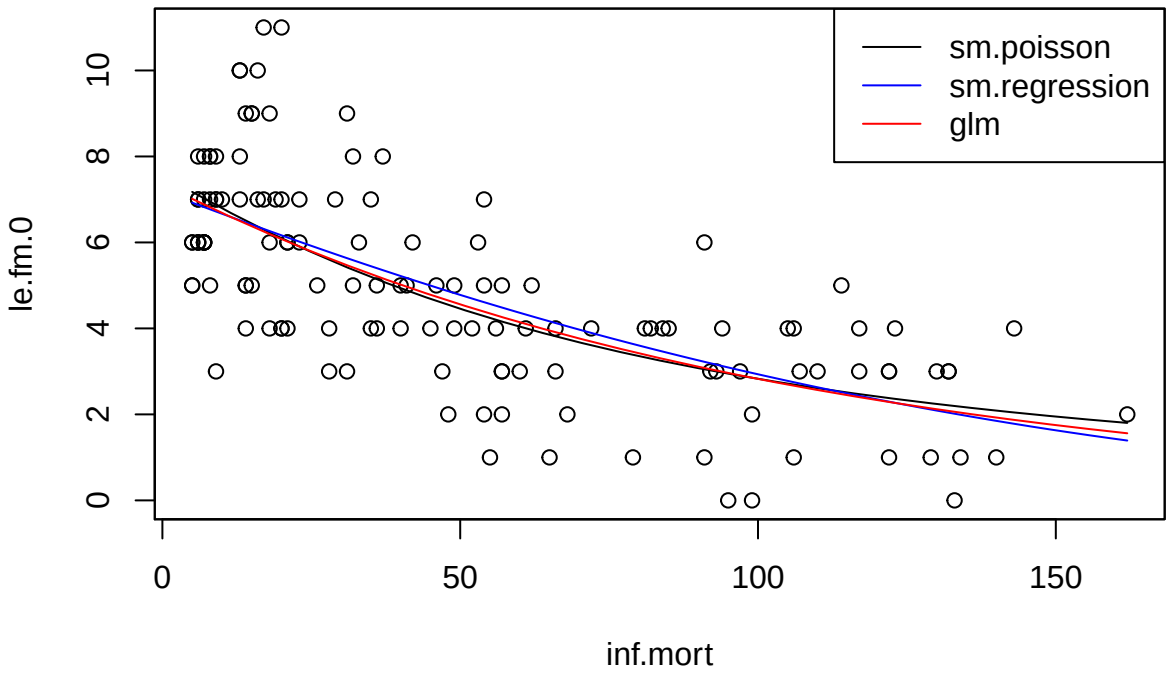
First, we are going to implement the function to choose the bandwith for local Poisson regression using log-likelihood with cross-validation method (loglik.CV). As a matter of practice, we will compare the bandwidth obtained with the probability of missclassification method with cross-validation (prob.missclas.CV)

```
# Write your own CV bandwidth choice script for the local Poisson regression.
# h.cv.sm.possion.R
#
# method can be equal to 'loglik.CV' (default) or 'prob.missclas.CV'
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"))#,hcv(x,y))
    rg.h <- range(hh)
  }
  i <- 0
  gr.h <- exp( seq(log(rg.h[1]/1.1), log(rg.h[2]*1.1), 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)]))
}
# method 1
prob.missclas.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")$estimate
                 },   x,y,h)
  return(sum(abs(pred-y)>.5)/n)
}
# method 2
loglik.CV <- function(x,y,h){
  n <- length(x)
  lambda <- sapply(1:n,
                 function(i,x,y,h){
                   sm.poisson(x=x[-i],y=y[-i],h=h,eval.points=x[i],display="none")$estimate
                 },   x,y,h)
  # using minus here is aim to keep using min later
  return (-sum(log(exp(-lambda)*(lambda^x)/(factorial(x)))))
}
```

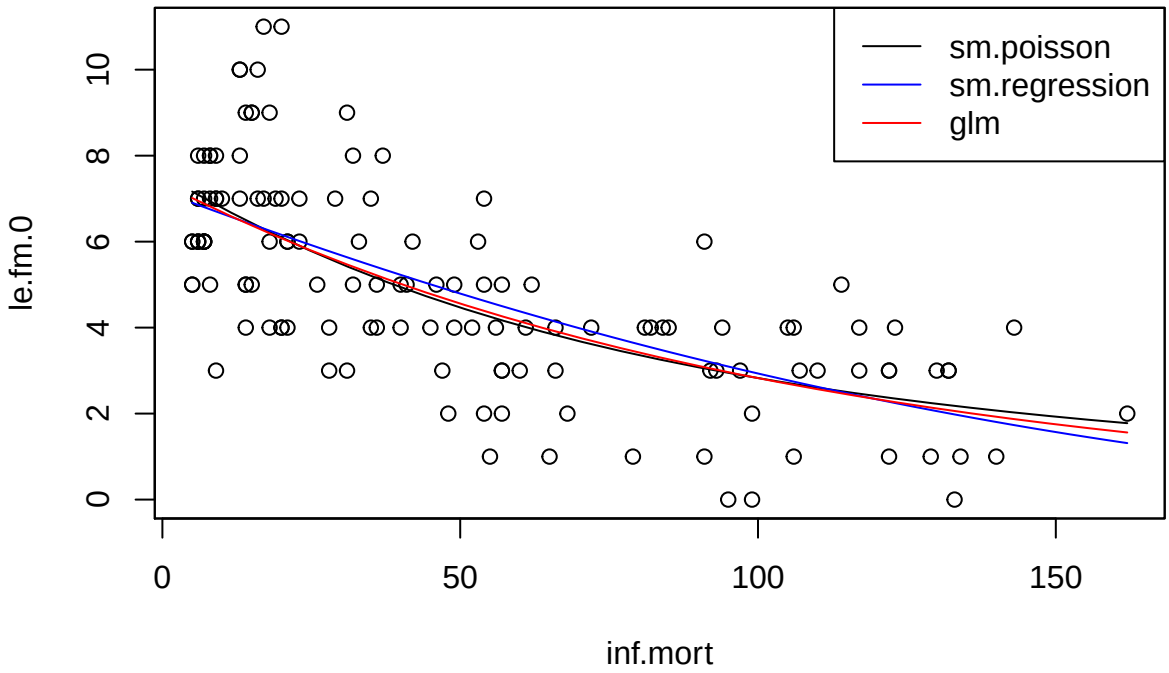
Now, we are going to use the previous function to choose the bandwidth and fit a model with le.fm.0 as response variable and infant mortality rate (inf.mort) as explanatory variable.

```
# Maximizing the log-likelihood of a new observation
h.list.lk <- h.cv.sm.poisson(inf.mort,le.fm.0,method = loglik.CV)

aux <- sm.poisson(inf.mort,le.fm.0,h=h.list.lk$h.cv, col=1)
sm.regression(inf.mort,le.fm.0,h=h.list.lk$h.cv,col=4,add=T)
```

```r
aux.glm <- glm(le.fm.0 ~ inf.mort,family=poisson)
pred <- predict(aux.glm,
                newdata=data.frame(inf.mort=aux$eval.points),
                type="response")
lines(aux$eval.points,pred,col=2)
legend('topright', c("sm.poisson", "sm.regression", "glm"),
       col = c('black', 'blue', 'red'), lty = c(1, 1, 1))
```



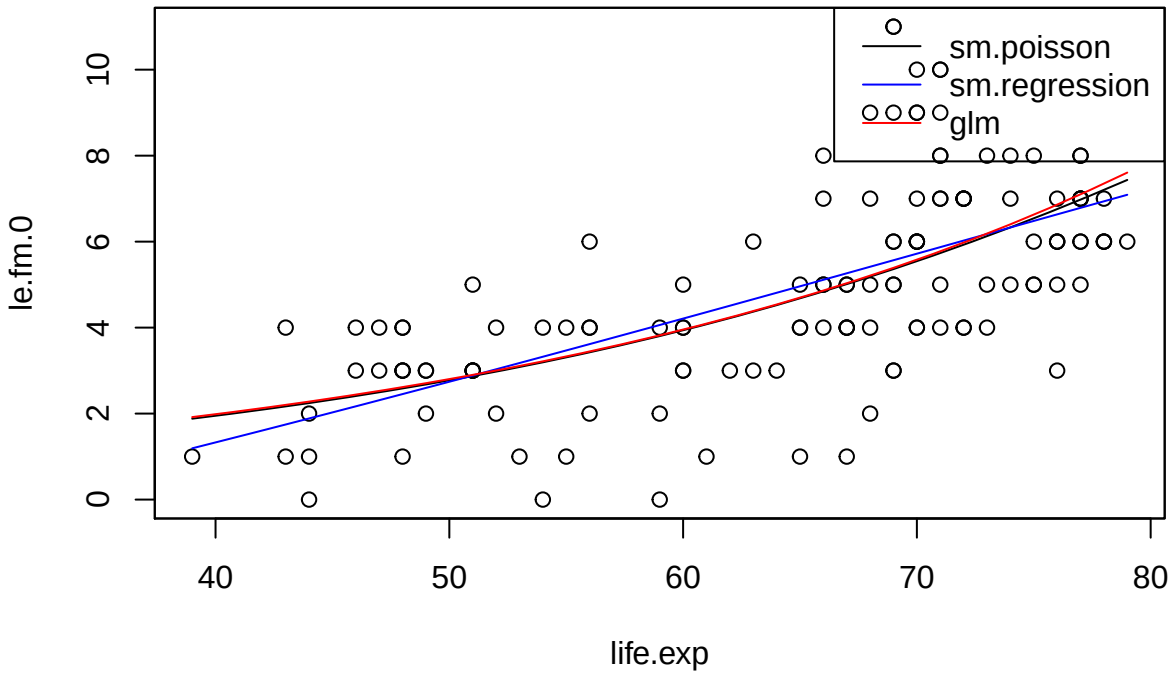

```r
# Minimizing the probability of misclassification of a new observation
h.list.pm <- h.cv.sm.poisson(inf.mort,le.fm.0,method = prob.missclas.CV)

aux <- sm.poisson(inf.mort,le.fm.0,h=h.list.pm$h.cv, col=1)
sm.regression(inf.mort,le.fm.0,h=h.list.pm$h.cv,col=4,add=T)
aux.glm <- glm(le.fm.0 ~ inf.mort,family=poisson)
pred <- predict(aux.glm,
                newdata=data.frame(inf.mort=aux$eval.points),
                type="response")
lines(aux$eval.points,pred,col=2)
legend('topright', c("sm.poisson", "sm.regression", "glm"), col = c('black', 'blue', 'red'), lty = c(1,
```



From the previous plots we can see that the three models show the same behaviour, meaning that the all three can explain the response variable equally good.

Then we fit another model using this time life expectancy (life.exp) as explanatory variable

```
h.list <- h.cv.sm.poisson(life.exp,le.fm.0,method = loglik.CV)#rg.h=c(4,11)

aux <- sm.poisson(life.exp,le.fm.0,h=h.list$h.cv, col=1) # black
sm.regression(life.exp,le.fm.0,h=h.list$h.cv,col=4,add=T) # blue
aux.glm <- glm(le.fm.0 ~ life.exp, family=poisson)
pred <- predict(aux.glm,
                newdata=data.frame(life.exp=aux$eval.points),
                type="response")
lines(aux$eval.points,pred,col=2)
legend('topright', c("sm.poisson", "sm.regression", "glm"),
       col = c('black', 'blue', 'red'), lty = c(1, 1, 1))
```



We can see that the parametric model fitted with a Poisson GLM (red line) and the nonparametric Poisson regression (sm.poisson) fit with h bandwidth calculated by function h.cv.sm.poisson match pretty well. Meanwhile, the standard nonparametric fit (blue line) is likely a straight line.