Task 4 : Bandwidth choice in the local Poisson model

Martin Guy and Hannes Leskela

December 17, 2016

Introduction

The goal of this exercise is to implement bandwidth choice functions for the local Poisson regression, and fit that model working with the file `countries.txt`. This file contains information on development indicators measured in 132 countries (Source: World Bank,1992). We will then compare our model with a standard nonparametric regression fit (with `sm.regression`) and a parametric fitting of a Poisson Generalized Linear Model (using `glm`).

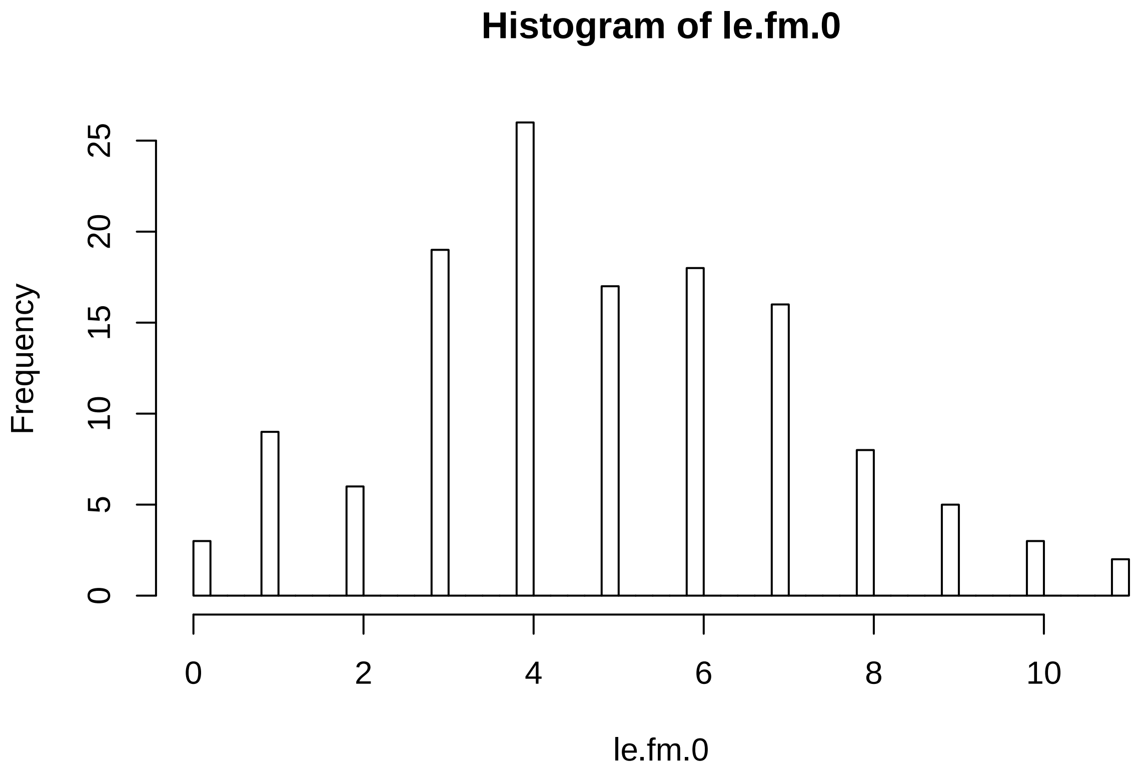
Analysing the data

First look

We will be working with the file `countries.txt` containing information on development indicators measured in 132 countries. We will focus on the following variables:

- `life.exp` (Life expectancy at birth)
- `inf.mort` (Infant mortality rate)
- `le.fm` (Difference Life expectancy at birth for females minus Life expectancy at birth for males)

The variable `le.fm` always takes non-negative values, except for one country, so we get rid of it and now consider `le.fm.0`. Here is a histogram of `le.fm.0`.



We can observe than in every country (except one) the life expectancy at birth for females is higher than for males.

Choice of the bandwidth

We modified the `h.cv.sm.binomial.R` file to a `h.cv.sm.poisson.R` file which calculates the bandwidth for a local Poisson regression. Using it we find a value for the bandwidth for our local Poisson model (`h1` for the first local Poisson model and `h2` for the second one):

```
## h1 = 55.0671548918819
```

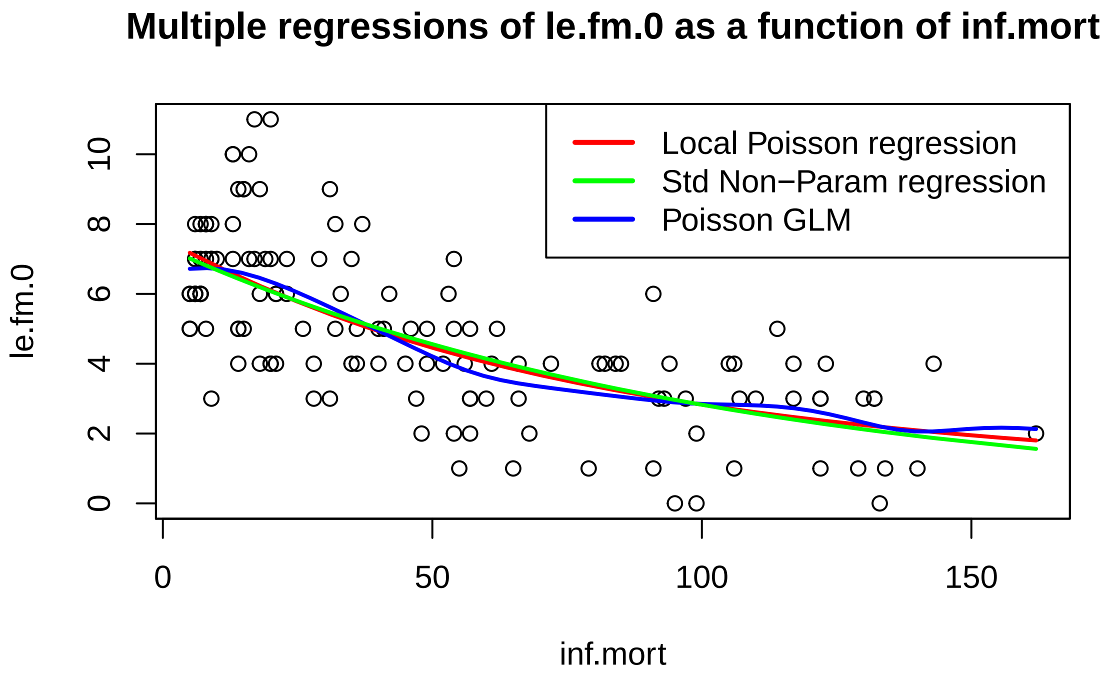
```
## h2 = 22.8456097023355
```

We select this bandwidth by cross-validation using the log-likelihood function of a Poisson distribution with a parameter λ:

$$l(\lambda, x) = \sum_{i=1}^{n} x_i log(\lambda) - n\lambda$$

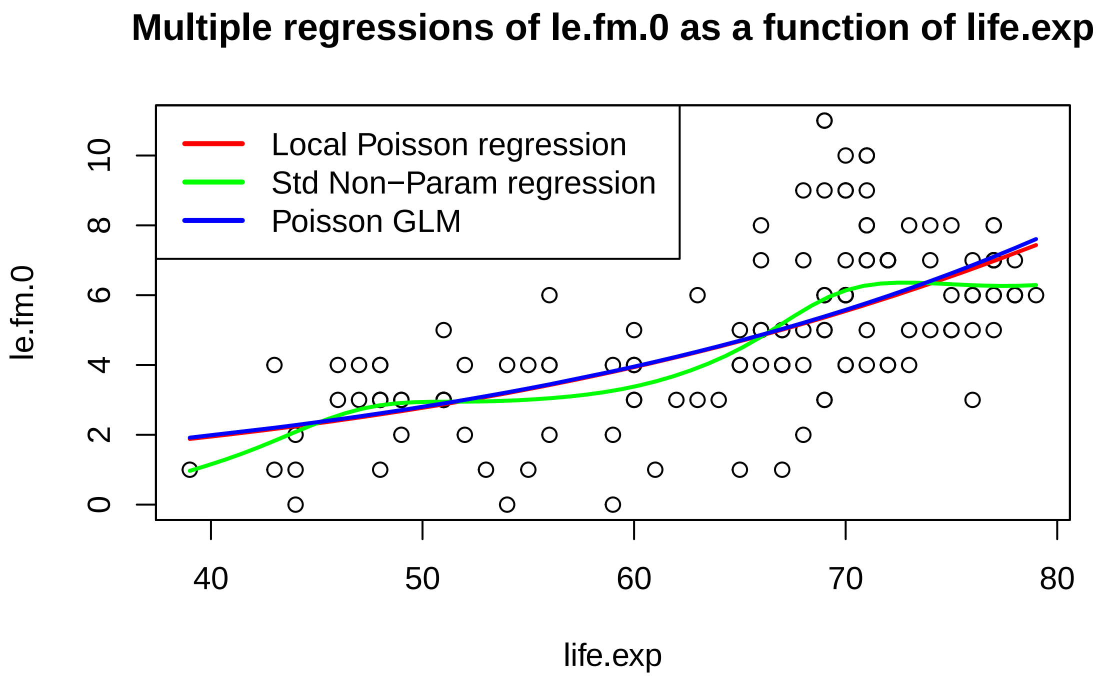
First models: le.fm.0 as a function of inf.mort

First, we will modelize `le.fm.0` as a function of infant mortality rate. In every case we will use local Poisson regression (red), a standard nonparametric regression fit (green), using a Poisson Generalized Linear Model (blue).



Second models: le.fm.0 as a function of life.exp

Then, we will modelize `le.fm.0` as a function of life expectancy at birth.



Conclusion

In each model, the optimal value of the bandwidth is selected. The models do not behave the same, which explains why they do not have the same optimal bandwidth. Nevertheless, the models seem to have quite similar fits.