Estimating the Conditional Variance by Local Linear Regression

Marcel Pons Cloquells

December 11, 2020

In this homework we are going to use the `Aircraft` data from the `sm` library, which consists of records on six characteristics of aircraft designs which appeared during the twentieth century. From this data, we are going to build a non parametric local linear regression using `Year` as the explanatory variable and the logarithm of `Weight` as response variable (maximum take-off weight in kg).

Two local linear regression models will be built, one using the created function `locpolreg` and the other using the `sm.regression` function from the `sm` library. For each model the optimal bandwith value h will be found and used for constructing the final models.

Finally, the conditional variance σ^2 will be estimated for each model.

1 Choosing the best bandwidth value

Two methodologies will be followed in order to find the optimal bandwidth value: *Leave-one-out cross validation* for the `logpolreg` model and *direct plug-in* for `sm.regression` (using `dpill` from `KernSmooth`).

For the first aforementioned methodology we use the built function `h.loocv`, which given the vectors x, y and a vector of candidate values of h it returns the MSPE of the local regression for each h. If we plot the MSPE value for each value of h, we can appreciate that the best bandwidth value h is 4.417

```
h.v <- exp(seq(from=log(0.3), to=log(15), length=17)) # considered candidates
h.loocv(x=Year, y=lgWeight, h.v=h.v)
```

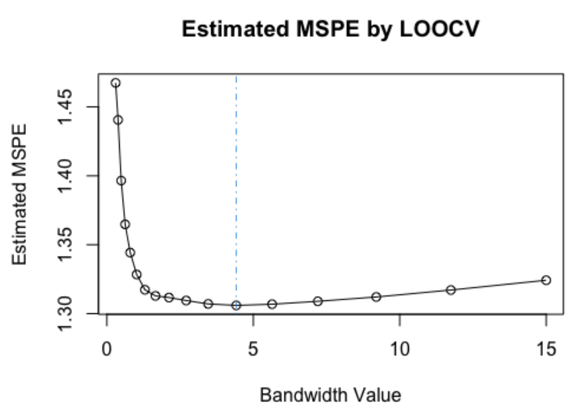


Figure 1: MSPE for the different bandwith values

For the other model, as stated before, we use the specific bandwidth selector for local regression: *direct plug-in*. With this method the best bandwidth value h is 5.433

```
h.dpi <- dpill(x=Year, y=lgWeight,gridsize = 101, range.x = range(Year))
```

2 Estimating the conditional variance

In order to estimate the conditional variance of `lgWeigth` given `Yr` for the two models, we apply the following procedure:

1. Fit a non-paremetric regression to data (x_i, y_i) and save the estimated values $\hat{m}(x_i)$.
2. Transform the estimated residuals $\hat{\epsilon} = y_i - \hat{m}(x_i) \rightarrow z_i = log\ \hat{\epsilon_i}^2 = log((y_i - \hat{m}(x_i))^2)$
3. Fit a nonparametric regression to data (x_i, z_i) and call the estimated function $\hat{q}(x)$. Observe that $\hat{q}(x)$ is an estimate of $log\ \sigma^2(x)$.
4. Estimate $\sigma^2(x)$ by $\sigma^2(x) = e^{\hat{q}(x)}$

And once we have the estimation, we plot $\hat{\epsilon}_i^2$ against x_i and superimpose the estimated function $\sigma^2(x)$ and also we plot the function $\hat{m}(x)$ and superimpose the bands $\hat{m}(x) \pm 1,96\hat{\sigma}(x)$.

`loc.pol.reg`

```
lpg.model <- locpolreg(x=Year,y=lgWeight,h=optimal_h,q=1,tg=Year)
m.hat <- lpg.model$mtgr

e_sq <- (lgWeight-m.hat)**2
z <- log(e_sq)

q <- locpolreg(x=Year,y=z,h=optimal_h,q=1,tg=Year)
sigma <- exp(q$mtgr)
```

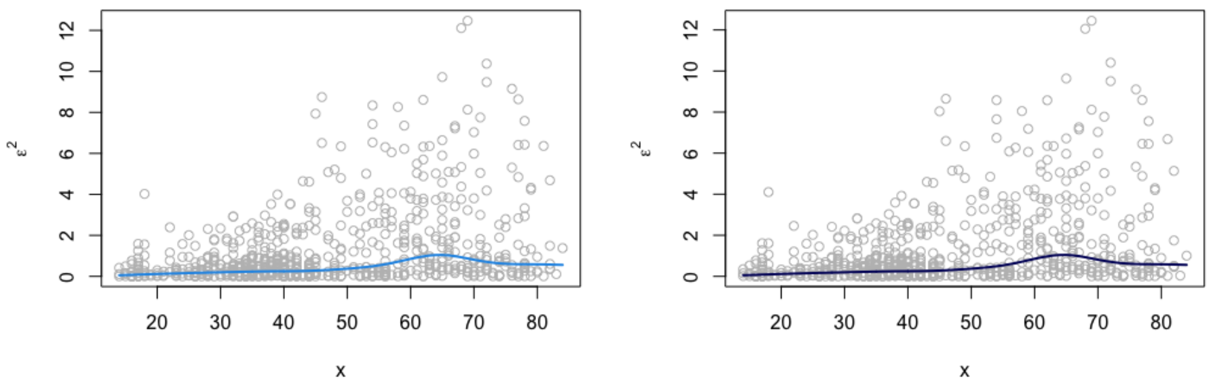
`sm.regression`

```
sm.lpr <- sm.regression(x=Year,y=lgWeight,h=h.dpi, eval.points=Year)
m.hat.sm <- sm.lpr$estimate

e_sq.sm <- (lgWeight - m.hat.sm)**2
z.sm <- log(e_sq.sm)

q.sm<- sm.regression(x=Year,y=z,h=h.dpi,eval.points=Year)
sigma.sm <- exp(q.sm$estimate)
```

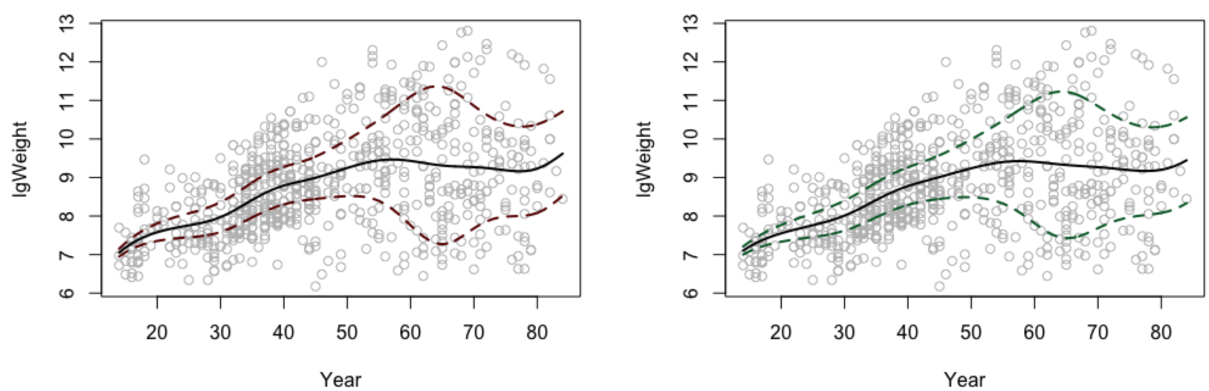
By comparison of the plots it can be appreciated that both functions of the linear local regression with the bandwidth value chosen with different techniques are very similar. From Figure 3 a little difference can be appreciated in the bands $\hat{m}(x) \pm 1,96\hat{\sigma}(x)$, which are slightly wider (almost inappreciable) when using a smaller h (4.417). This fact makes sense because the smaller is the h value, the more flexible is the model.



(a) `logpolreg` with $h = 4.417$

(b) `smregression` with $h = 5.433$

Figure 2: Plot of $\hat{\epsilon}_i^2$ against x_i with the estimated function $\sigma^2(x)$ superimposed.



(a) `logpolreg` with $h = 4.417$

(b) `smregression` with $h = 5.433$

Figure 3: Plot of the function $\hat{m}(x)$ with the bands $\hat{m}(x) \pm 1,96\hat{\sigma}(x)$ superimposed.