Estimating the conditional variance by local linear regression

Joel Cantero Priego & Ricard Meyerhofer Parra

22/11/2019

Introduction

In this assignment, we are going to use the Aircraft data from the R library sm. These data record six characteristics of aircraft designs which appeared during the twentieth century.

Variable name	Description	Values
Yr	year of first manufacture	Integer
Period	a code to indicate one of three broad time periods	Integer
Power	total engine power (kW)	Integer
Span	wing span (m)	Integer
Length	length (m)	Integer
Weight	maximum take-off weight (kg)	Integer
Speed	maximum speed (km/h)	Integer
Range	range (km)	Integer

We transform data taken logs (except Yr and Period): lgPower, . . . , lgRange.

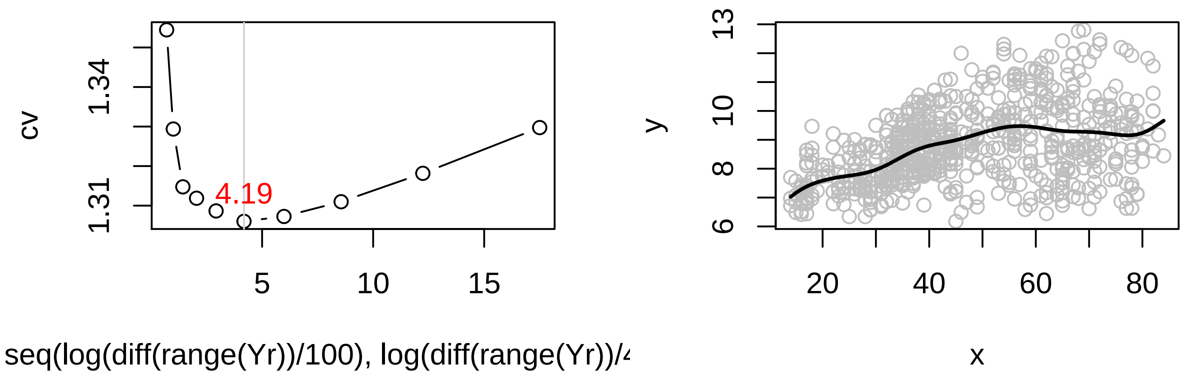
The main objective of this project is to estimate the coniditional variance of *lgWeight* (variable Y) given Yr (variable x) using two different procedures:

- **loc.pol.reg** function that we can find in ATENEA choosing all the bandwith values we need by leave-one-out cross-validation.
- **sm.regression** from library sm choosing all the bandiwth values we need by direct plug-in (using the function dpill from the same library KernSmooth).

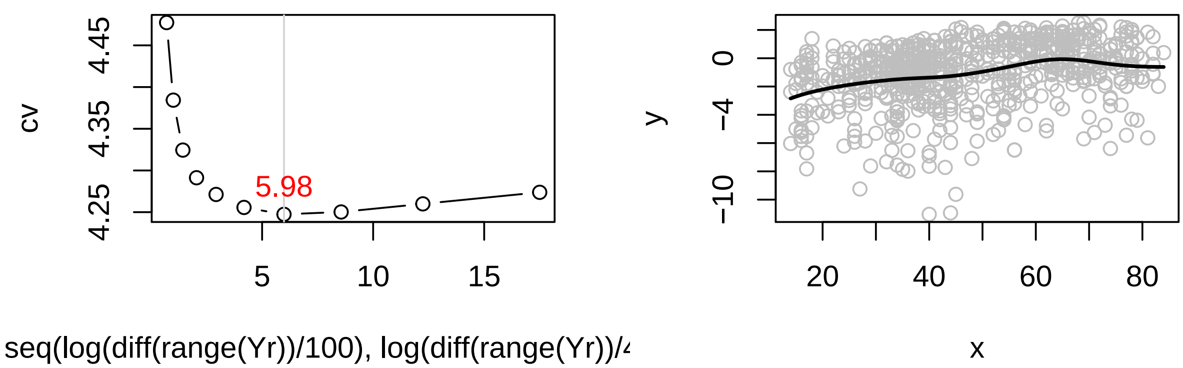
Using *locpolreg* function

Thanks to **loc.pol.reg function** from **locpolreg.R script** and **h.cv.gcv** and **h.k.fold.cv** from **Bandwidth_choice.Rmd** we are going to choose the **bandwith by leave-one-out cross-validation** using the **Gaussian method** (normal).

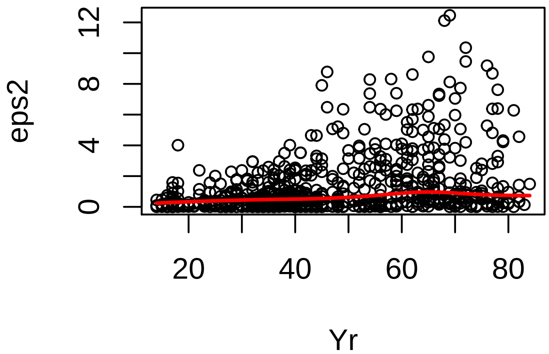
First of all, we define the bandwith candidates and we select the minimum one (4.19). Then, we can perform the local linear regression thanks to **locpolreg function** for the response variable **lgWeight** depending on the explanatory variable **Yr**. We obtain as well the residual values.

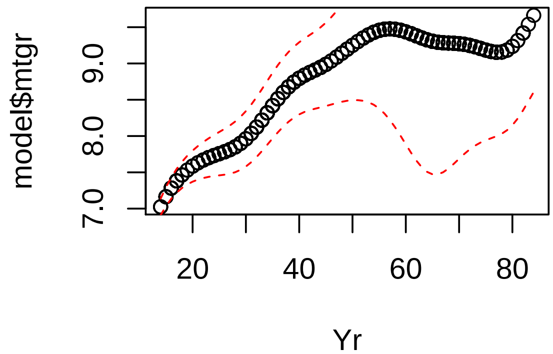


Once we have obtained the residual values $\hat{\epsilon}_i$ we transform the estimated residuals to $z_i = \log \hat{\epsilon}_i^2$. Finally, we fit a nonparametric regression to data (x_i, z_i) and call the estimated function $\hat{q}(x)$, that is an estimate of $\log \sigma^2(x)$. We perform a new model with a new bandwidth



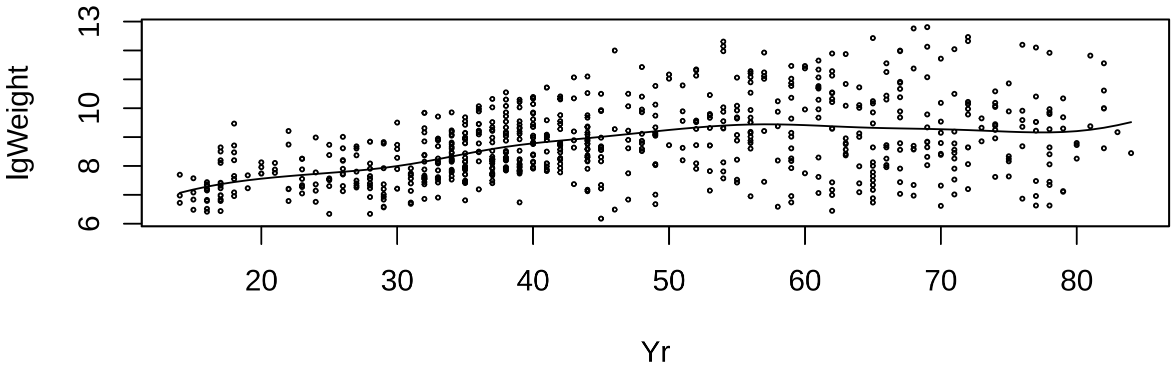
Finally, we draw a graphic of ϵ_i^2 against x_i and superimpose the estimated function $\hat{\sigma}^2(x)$. Lastly we draw the function $\hat{m}(x)$ and superimpose the bands $\hat{m}(x) \pm 1.96\hat{\sigma}(x)$.



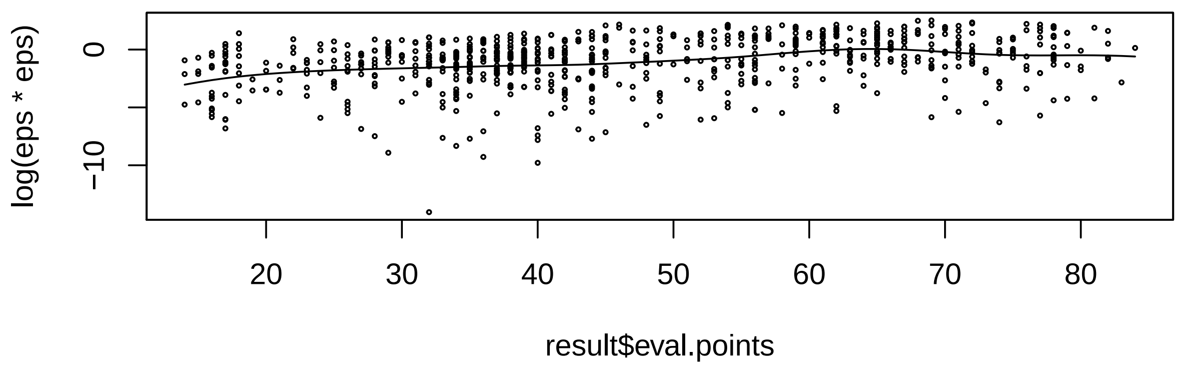


Using *sm.regression* function

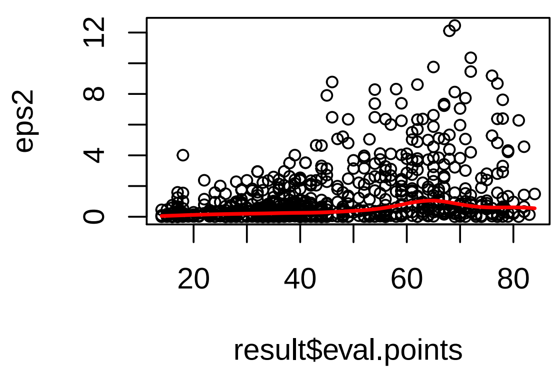
In this second part, we will follow the same steps as the previous part but now using **dpill** function from **KernSmooth** package to get the **Plug-in** parameter. Then, we compute the **sm.regression** function from the **sm package** with this bandwidth as parameter. We will obtain the local linear regression models $\hat{m}(x)$ and $\hat{q}(x)$. As before, we select the Gaussian kernel and we compute the residual values.

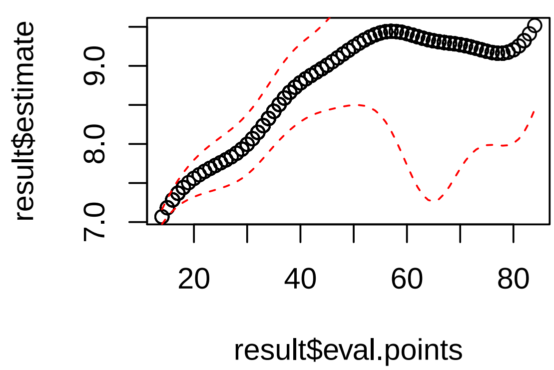


Once we have performed the local linear regression we can compute the residual values $\hat{\epsilon}_i =$ and z_i to generate a new model for z_i against x_i.



$\hat{\sigma}^2(x) = \exp \hat{q}(x)$ is the conditional variance where $\hat{q}(x)$ is the estimate we have obtained from the previous model. Finally, we draw a graphic of ϵ_i^2 against x_i and superimpose the estimated function $\hat{\sigma}^2(x)$. Lastly we draw the function $\hat{m}(x)$ and superimpose the bands $\hat{m}(x) \pm 1.96\hat{\sigma}(x)$.

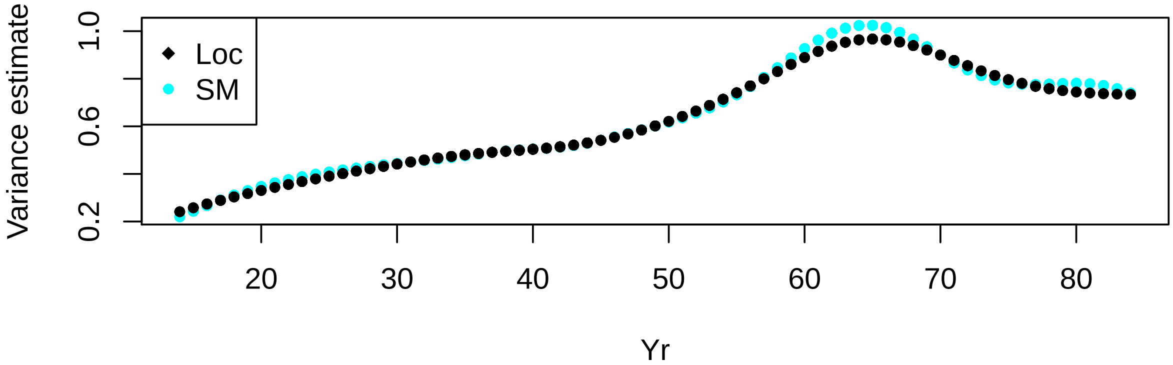




Conclusion

We can say that the bandwithvalues obtained from each method are close. In the case of LocPolReg, the first model obtained value is **4.18** and for the second one is **5.98**. In Sm.Regression case, the first bandwith model we obtain **5.02** and for the second model **4.28**.

If we plot these values, we can see that the shape is more or less similar but Sm.Regression is a little bit more extrem than LocPolReg.



Loc
SM
Variance estimate
0.2
0.6
1.0
20
30
40
50
60
70
80
Yr