Estimating the conditional variance by local linear regression

Mattia Barbero - David Cardoner - Arnau Mercader

Objective

From one side, fit a non parametric regression with two different approaches to obtain the optimal bandwith, firstly with leave one out cross-validation and secondly with plug-in method implemented in kernelSmooth package. On the other side, estimate the conditional variance $\sigma^2(x_i)$ with the two methods described above.

In the next chunk we read the aircraft dataset and perform transformations of variables applying logarithm.

```r
library(sm)
sm.options(display="none") #to avoid generate the plot.
library(dplyr)
data(aircraft)
#help(aircraft)
aircraft<-aircraft %>%
  mutate_at(.vars = vars(Power,Span,Length,Weight,Speed,Range),
            .funs = funs(log))
noms <-unlist(strsplit(paste0("log",names(aircraft)[3:length(names(aircraft))],
                              collapse = ","),split = ","))
noms <- c("Yr","Period",noms)
names(aircraft)<-noms
```

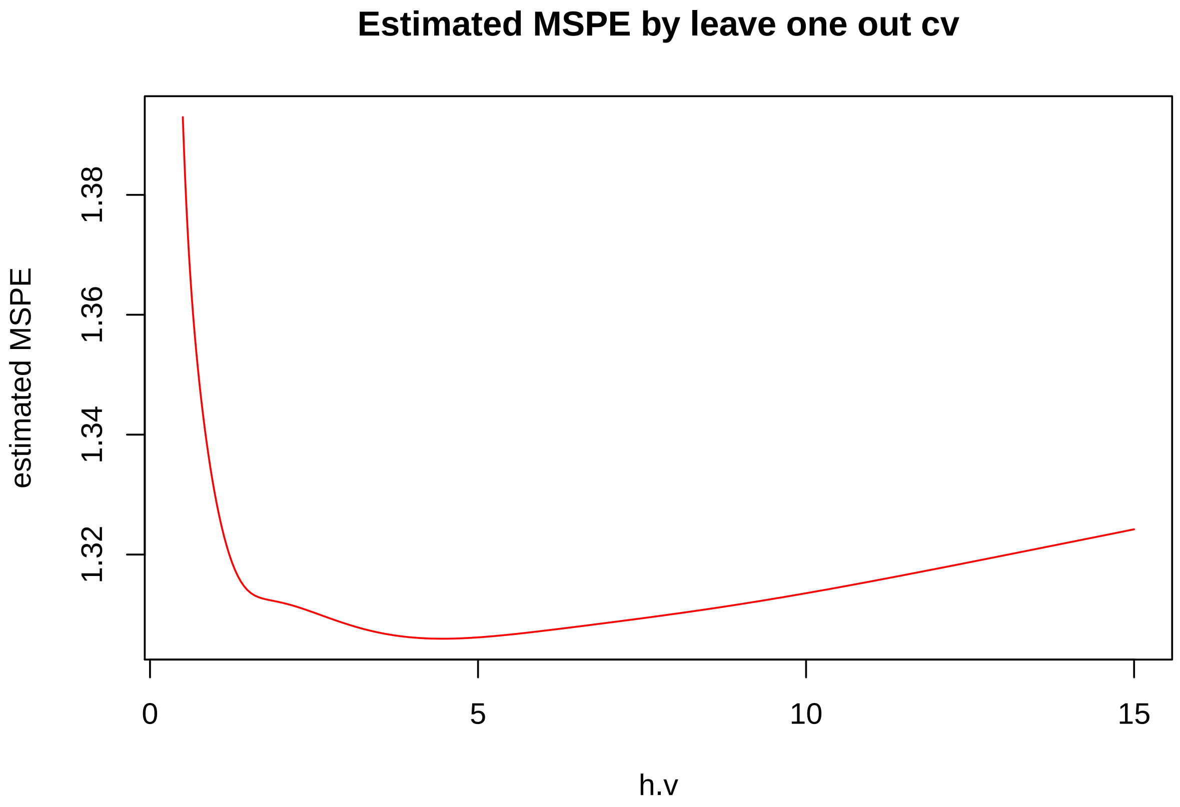
First approach: Leave one out cross-validation with locpolreg

We call the functions defined in Atenea. For implement leave one out cross-validation we evaluate the function k-fold equal to n, so we are using all the observations in the dataset.

```r
source("locpolreg.r")
source("k_fold_cross_oneleave_smooth.r")
```

With locpolreg function and the modified k_fold cross-validation we find the optimal **h (bandwidth)** evaluated in a grid of 101 points between $log(0.5), ..., log(15)$. In the next plot we can see the optimal bandwidth comparing the estimated **MSPE** with different bandwidths.

```r
##Provar de posar altres Kernels no nomes Normal.
h.v <-  exp( seq(from=log(.5), to = log(15), length=101))
out.lout.cv <- h.k.fold.cv(x=aircraft$Yr, y=aircraft$logWeight,
                           h.v=h.v, k=length(aircraft$Period))
##con esto se hace un kernel gaussiano (Normal). Provar epanechnikov.
y.max <- max(out.lout.cv$k.cv)
y.min <- min(out.lout.cv$k.cv)

plot(h.v,out.lout.cv$k.cv,ylim=c(y.min,y.max),ylab="estimated MSPE",type = "l",col=2,
     main="Estimated MSPE by leave one out cv")
```

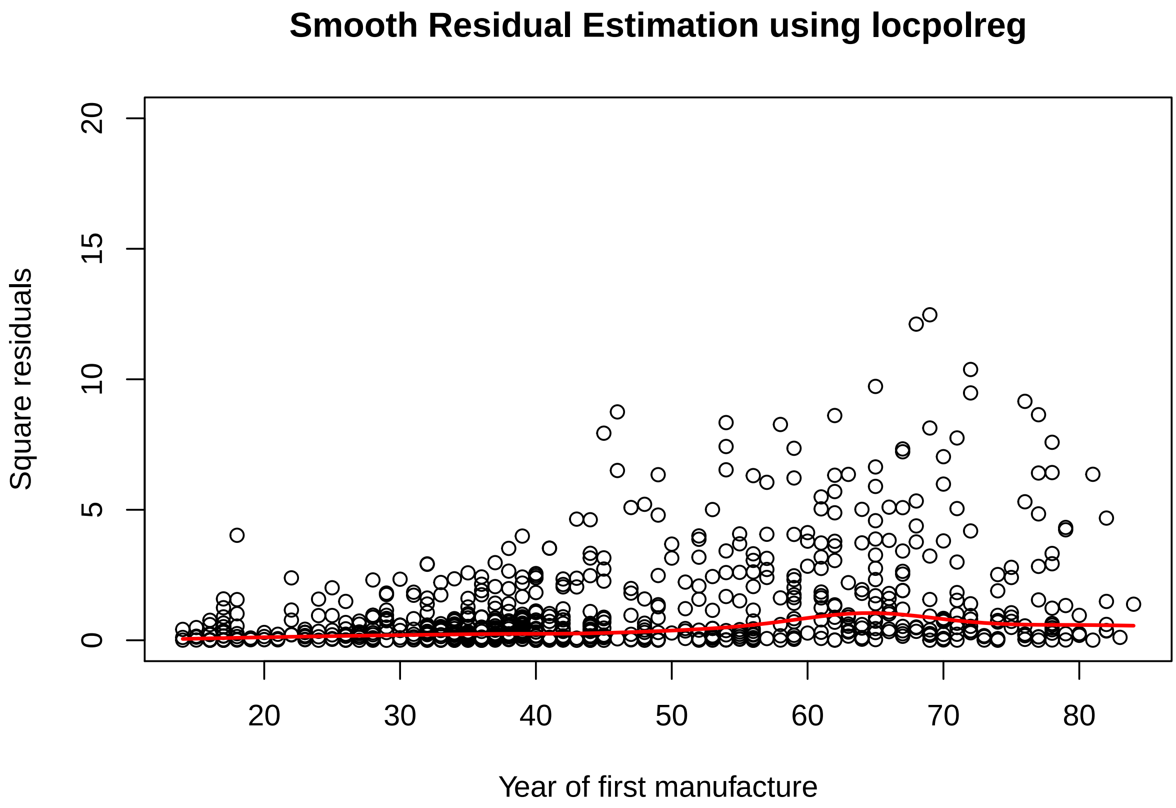


```
##valor optimo de h (o bandwith).
h.out<-out.lout.cv$h.v[which.min(out.lout.cv$k.cv)]
```

The best value = 4.409. Now we use the locpolreg functon with the best bandwidth to estimate the logarithm maximum take-off weight (kg) versus the years of manufacture. The type of kernel considered is a normal kernel. Once obtained the $\hat{m}(x)$ function now we calculate the residuals and perform $z_i = log(r_i^2)$, where r_i denote the residuals of model. With these z_i variables, we reuse locpolreg function with the same bandwidth (we could discuss if we should change the bandwidth or not) and same kernel. With the results of this estimation we perform $\sigma^2(x_i) = exp(\hat{q_x})$, where $\hat{q_x}$ are the estimations of the previous model (evaluated with locpolreg function).

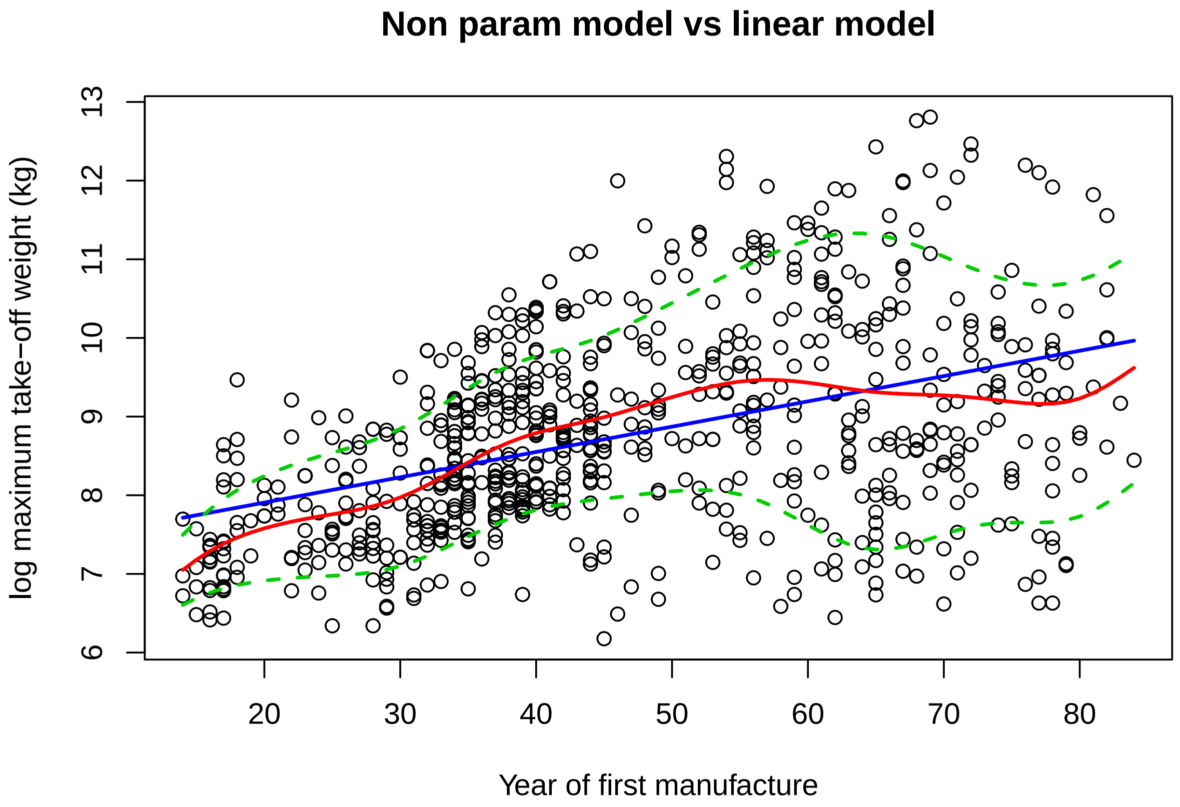
```
mx <- locpolreg(aircraft$Yr,aircraft$logWeight,h=h.out,type.kernel="normal",doing.plot=FALSE)
mxpred <- mx$mtgr
residuals <- (aircraft$logWeight-mxpred)
zi <- log(residuals^2)

qx <- locpolreg(aircraft$Yr,zi,h=h.out,type.kernel="normal",doing.plot=FALSE)
sigma2x<-exp(qx$mtgr)
```

```
plot(aircraft$Yr,residuals^2,ylim = range(0,20),
xlab="Year of first manufacture",
ylab = "Square residuals", main = "Smooth Residual Estimation using locpolreg")
lines(aircraft$Yr,sigma2x,col=2,lwd=2)
```



Now, we plot with points the log of the maximum take-off weigth (in Kgs) vs the years of manifacture. Our non-parametric regression function (red line) and a linear regression function (blue line) is also drawn. As shown, the value of the non-parametric function is lower than the value of the linear regression with low and high values of the years of construction.In addition, there are represented the confidence bands (green-dashed lines), given by a value of 1.96 the variance more (or less) the predicted value.

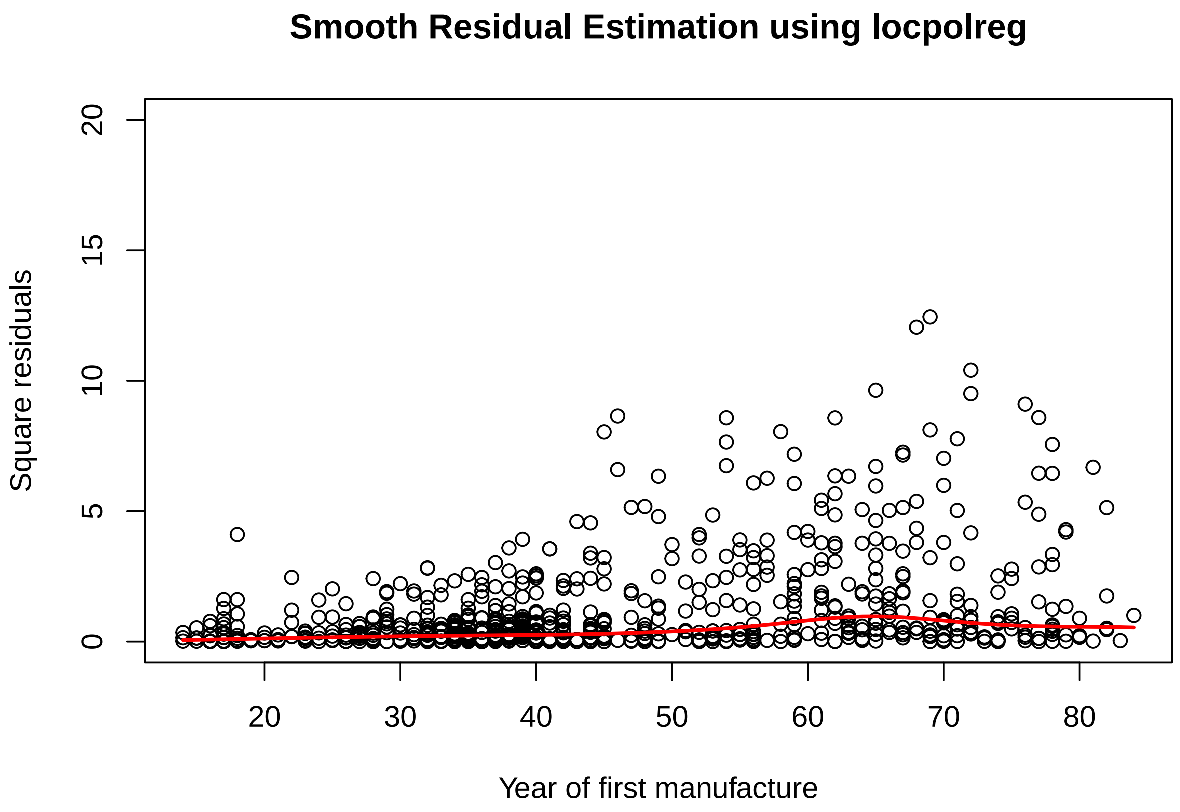
```
plot(aircraft$Yr,aircraft$logWeight,
xlab = "Year of first manufacture",ylab = "log maximum take-off weight (kg)",
main="Non param model vs linear model")
lines(aircraft$Yr,predict(lm(logWeight~Yr,data=aircraft)),col=4,lwd=2)
lines(aircraft$Yr,mxpred,col=2,lwd=2)
lines(aircraft$Yr,mxpred+1.96*(sqrt(sigma2x)),col=3,lwd=2,lty=2)
lines(aircraft$Yr,mxpred-1.96*(sqrt(sigma2x)),col=3,lwd=2,lty=2)
```



Now we use the dpill functon in order to estimate the optimal bandwidth of the kernel. The type of kernel considered is a normal kernel. Once obtained the $\hat{m}(x)$ is possible to calculate the residuals and perform $z_i = log(r_i^2)$, where r_i denote the residuals of model. With these z_i variables, we use the sm.regression function with the bandwidth calculated before. With the results of this estimation we perform $\sigma^2(x_i) = exp(\hat{q_x})$, where $\hat{q_x}$ are the estimations of the previous model.

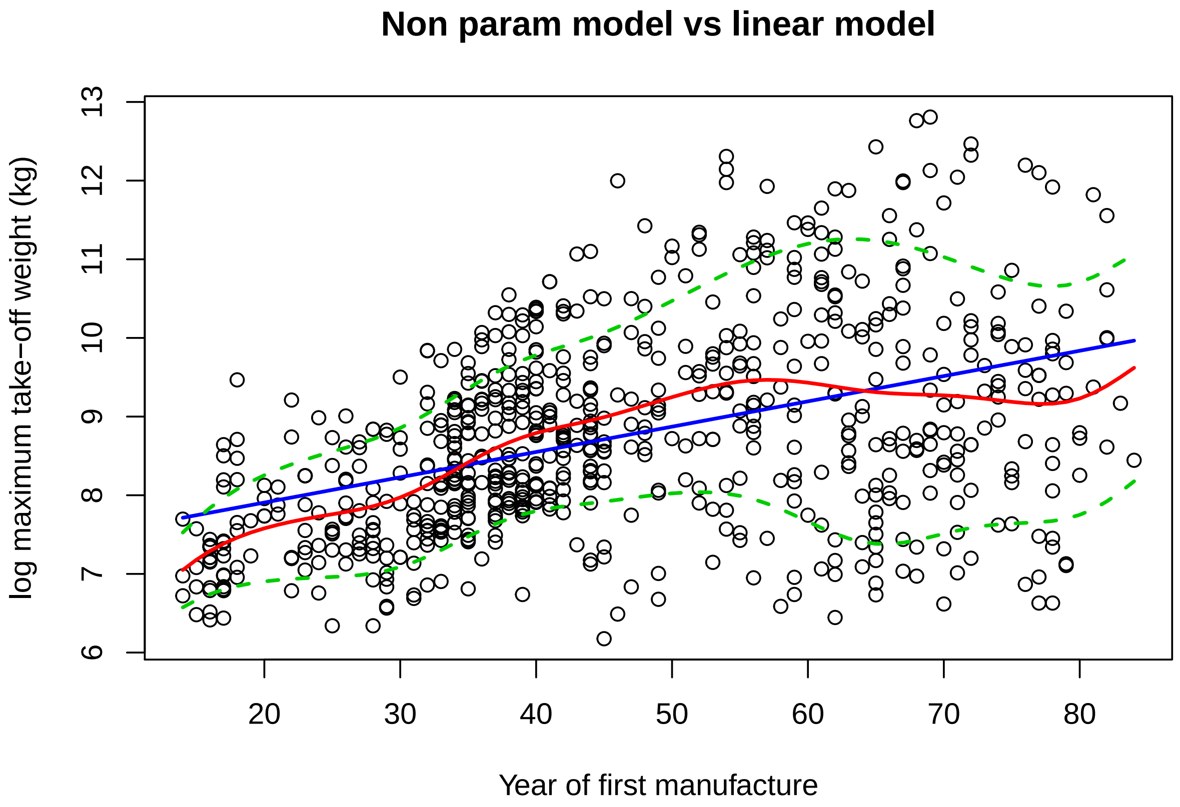
Second approach: Plug-in

```
library(sm)
library(KernSmooth)
h.dpill <- dpill(x=aircraft$Yr,y=aircraft$logWeight,gridsize=101,
                 range.x=range(aircraft$Yr))

smpred <- sm.regression(x=aircraft$Yr,y=aircraft$logWeight,h=h.dpill,
eval.points=aircraft$Yr)

mxpredsm <- smpred$estimate
residuals <- (aircraft$logWeight-mxpredsm)
zi <- log(residuals^2)

qx <- sm.regression(x=aircraft$Yr,y=zi,h=h.dpill,eval.points=aircraft$Yr)
linealm<-lm(logWeight~Yr,data=aircraft)
sigma2x<-exp(qx$estimate)
plot(aircraft$Yr,residuals^2,ylim = range(0,20),
xlab="Year of first manufacture",
ylab = "Square residuals", main = "Smooth Residual Estimation using locpolreg")
# lines(aircraft$Yr,(linealm$residuals^2),col=4,lwd=2)
lines(aircraft$Yr,sigma2x,col=2,lwd=2)
```



Now we can draw the non-parametric regression function (red line) and the linear regression function (blue line), as before. Also we drawn the confidence band (green-dashed line). We can appreciate that this figure is very similar to that drawn using the locpolreg function.

```
plot(aircraft$Yr,aircraft$logWeight,
xlab = "Year of first manufacture",ylab = "log maximum take-off weight (kg)",
main="Non param model vs linear model")
lines(aircraft$Yr,predict(lm(logWeight~Yr,data=aircraft)),col=4,lwd=2)
lines(aircraft$Yr,mxpred,col=2,lwd=2)
lines(aircraft$Yr,mxpred+1.96*(sqrt(sigma2x)),col=3,lwd=2,lty=2)
lines(aircraft$Yr,mxpred-1.96*(sqrt(sigma2x)),col=3,lwd=2,lty=2)
```



Comparison MSE

Just to verify the results, we calculate the mean square error using the two smoothing functions and is proved that they are very similar. The difference is lower than the order of 0.01 (about 0.005).

```
mse <- function(a,b){mean((a-b)^2)}
mse_plugin <- mse(mxpredsm,aircraft$logWeight)
mse_l1out_cv <- mse(mxpred,aircraft$logWeight)
cat("mse obtained by plugin is:",mse_plugin,"\n")
```

```
mse obtained by plugin is: 1.281716
```

```
cat("mse obtained by leave one-out cv is:",mse_l1out_cv,"\n")
```

```
mse obtained by leave one-out cv is: 1.276712
```

```
cat("Difference",abs(mse_plugin-mse_l1out_cv),"\n")
```

```
Difference 0.00500384
```