Non-parametric models - Estimating conditional variance

Haoran Mo, Alexandra Yamaui

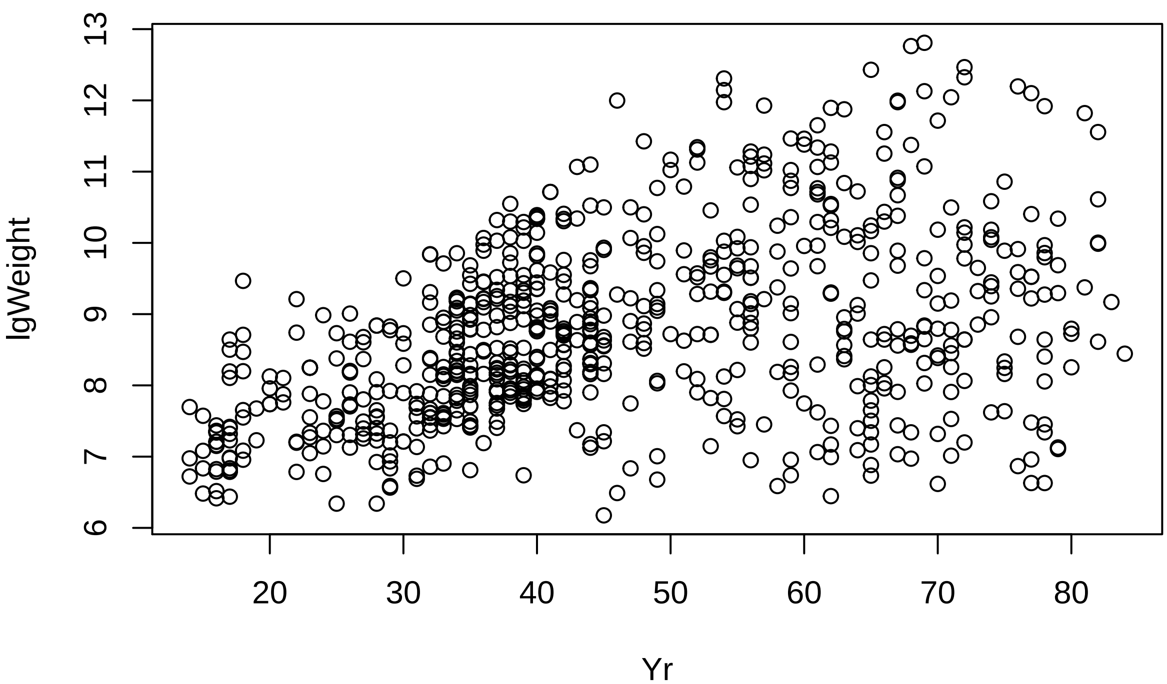
Having the heteroscedastic regression model

$$Y = m(x) + \sigma(x)\varepsilon = m(x) + \epsilon,$$

, where $E(\varepsilon) = 0, V(\varepsilon) = 1$

we want to estimate the function that represents the conditional variance σ^2 of the variable lgWeigth (log(Weight)) from the aircraft dataset, given that the explanatory variable (Yr) is equal to a value x. We will use nonparametric methods to do that. Below is the plot the response variable vs the explanatory variable.

```
data(aircraft)
attach(aircraft)
lgWeight <-log(Weight)
plot(Yr,lgWeight)
```



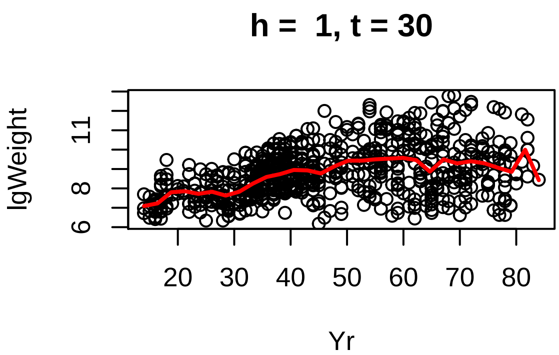
Initially, we are going to fit a local linear regression model to obtain an estimation $\hat{m}(x)$ of every point x. The general idea is to build a grid of intervals (t_i) centered around each point x and estimate a local linear regression in each interval. Doing this, we are going to try multiple sizes for the intervals and several values for the smoothing parameter h, which controls weight concentration around each point x.

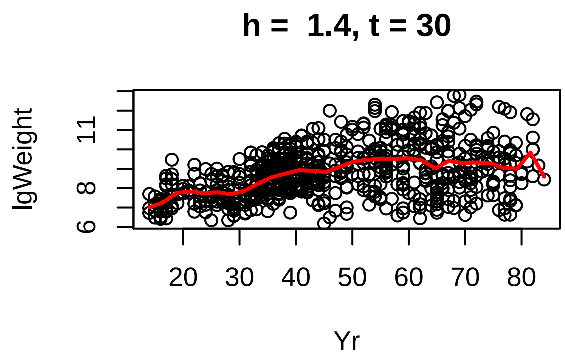
To make the regression function smooth weights are assigned to each pair (t_i, y_i) using a kernel function, which, in this case, is the Normal density function centered at 0, with h as standard deviation.

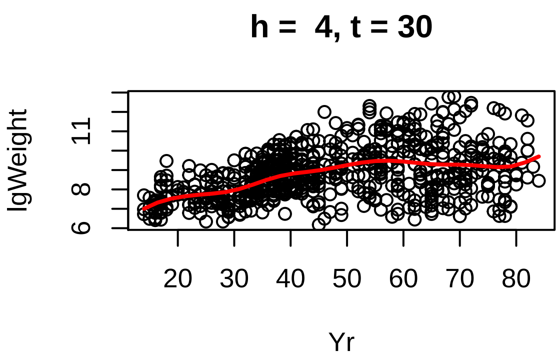
The result is shown below, where we can see an smooth function (in red) of the estimates.

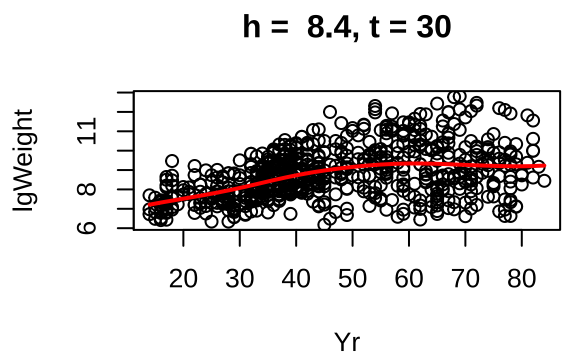
```
# step 1 Fit a nonparametric regression to data (xi,yi) and save
#   the estimated values m^ (xi).
par(mfrow=c(2,2))
```

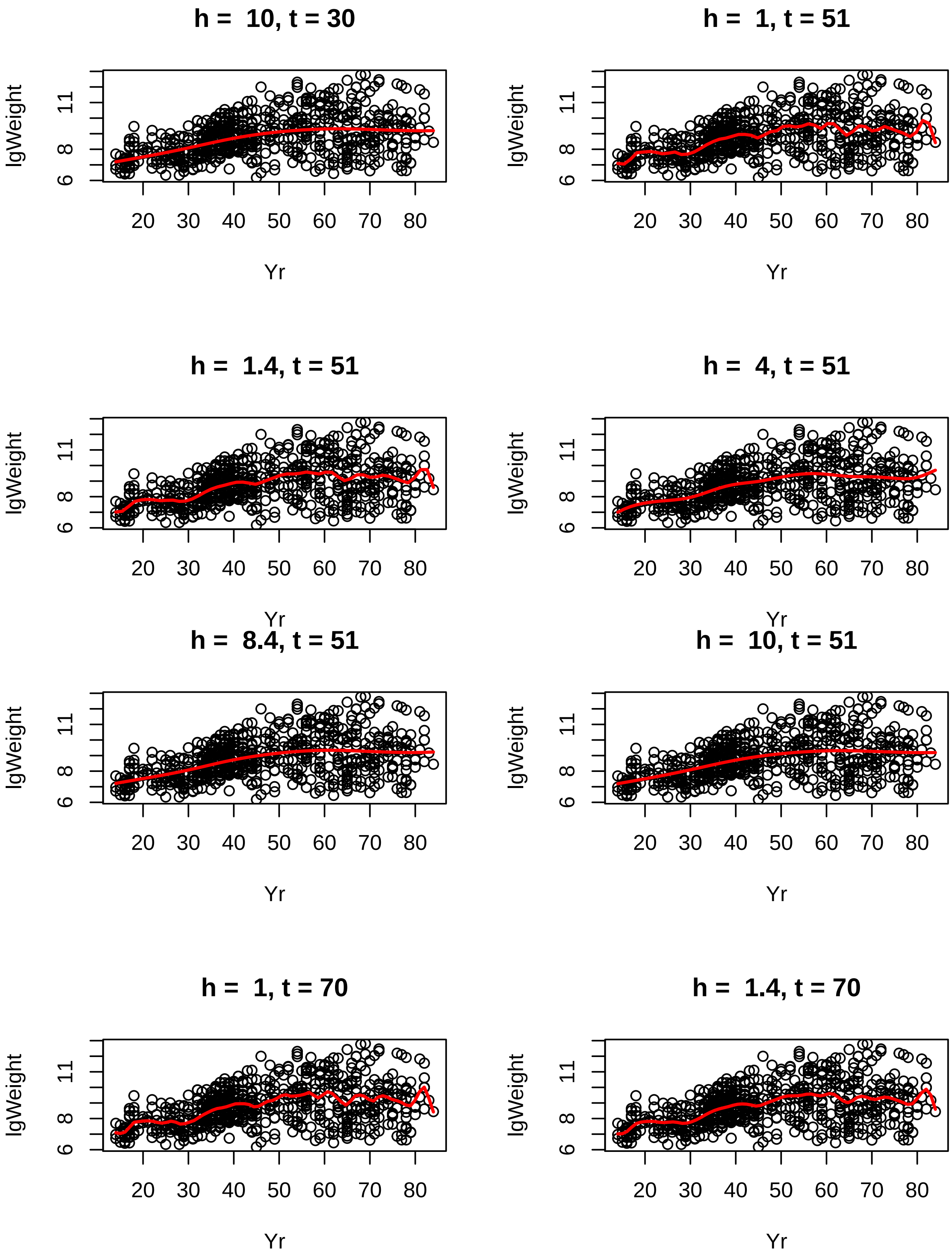
```r
hs <- c(1,1.4,4,8.4,10)
ts <- c(30,51,70)
for (t in ts) {
  for (h in hs) {
    tg = seq(min(Yr), max(Yr), length=t)
    llr <- loc.lin.reg(x=Yr, y=lgWeight, h=h, tg=tg)
    plot(Yr,lgWeight, main = paste("h = ", paste(h, paste(", t = ", t, sep = ""), sep = "")))
    lines(tg, llr$mt, col=2, lwd=2)
  }
}
```



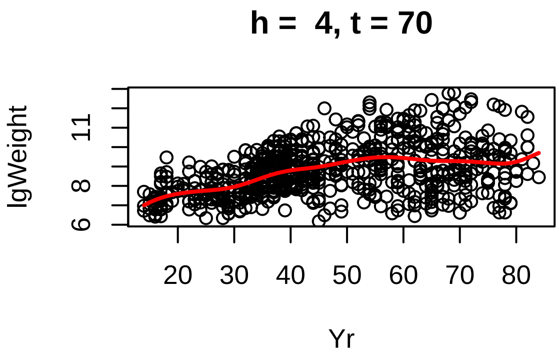


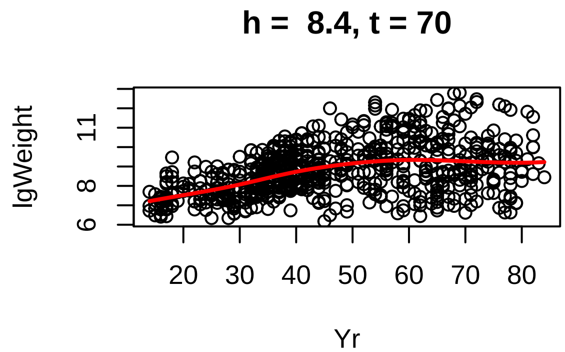


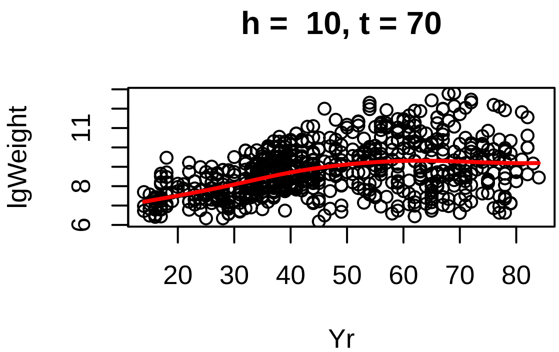





```
# TODO is this regression good enough?
```







Changing the smooth parameter and the size of the interval we can see that h has more impact than sizds of interval, as h increases the function curve becomes smoother. We will choose the combination $h = 4$ and $t = 70$

```
h <- 4
tg <- seq(min(Yr), max(Yr), length=70)
```

Now we are going to transform the estimated residuals applying logarithm to the square of it $\hat{\epsilon}_i = (y_i - \hat{m}(x_i))^2$, which represents the variance (the square deviation) of the model.

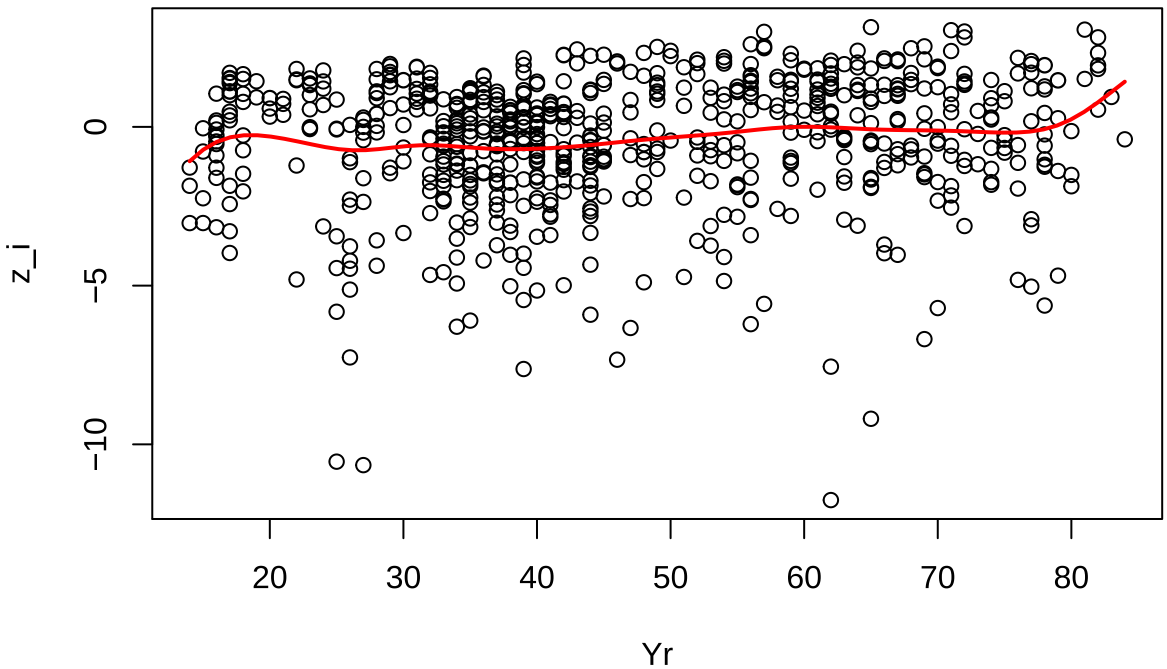
$$z_i = log\epsilon_i^2 = log((y_i - \hat{m}(x_i))^2)$$

```
# step 2 Transform the estimated residuals hat.e? = y_i - llr$mt
z_i = log((lgWeight - llr$mt)^2)
```

```
## Warning in lgWeight - llr$mt: longer object length is not a multiple of
## shorter object length
```

Then we perform a nonparametric regression over (x_i, z_i) to obtain the estimation of the (logarithm of the) variance $log\sigma^2(x)$.

```
# step 3 Fit a nonparametric regression to data (Yr,z_i) and
# call the estimated function q^(x).
llr2 <-loc.lin.reg(x=Yr, y=z_i, h=h, tg=tg)
plot(Yr,z_i)
lines(tg, llr2$mt, col=2, lwd=2)
```



Finally, we can obtain the conditional variance applying exponential to the estimation obtained before

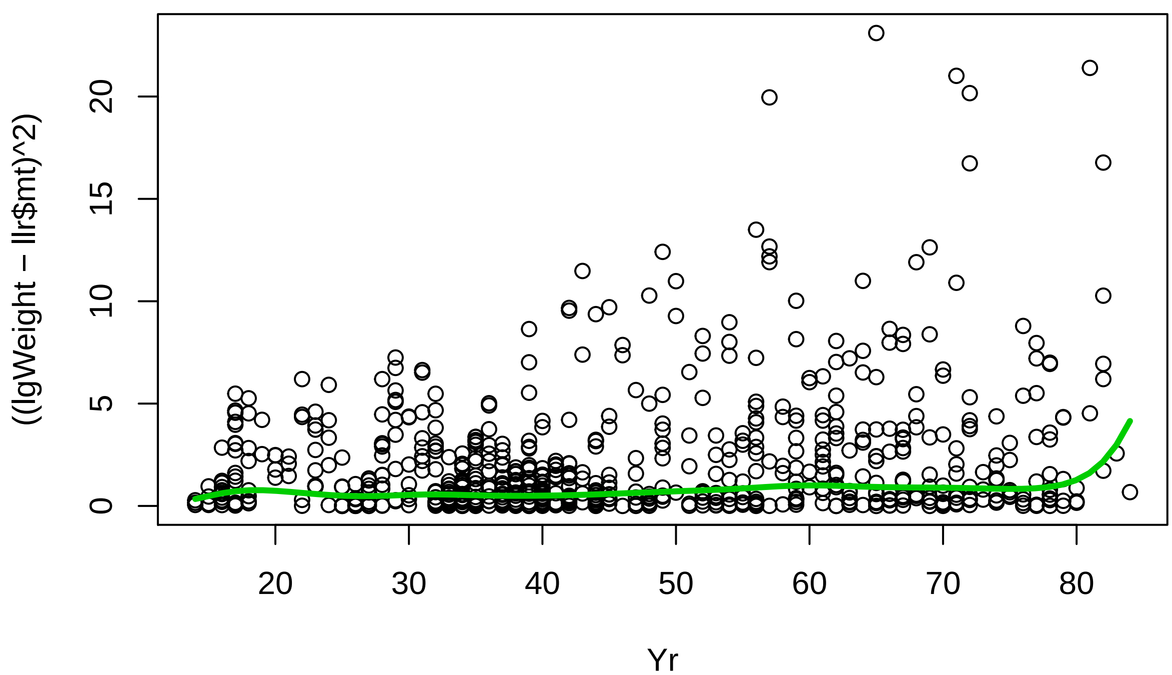
```
# step 4 Estimate sigma_2(x)
sigma_2 =  exp(llr2$mt)
```

Plotting the residuals against x_i we superimposed the estimated function $\hat{\sigma}^2(x)$

```
plot(Yr,((lgWeight - llr$mt)^2)) # residual square against x_i
```

```
## Warning in lgWeight - llr$mt: longer object length is not a multiple of
## shorter object length
```

```
lines(tg, sigma_2, col=3, lwd=3)
```



Here we present the draw of the function $\hat{m}(x)$ and the superimposed bands $\hat{m}(x) \pm 1.96\hat{\sigma}(x)$

```
sigma = sqrt(sigma_2)
plot(Yr,lgWeight)
lines(tg,llr$mt,col=3,lwd=3) # mt from step 1
lines(tg,llr$mt+1.96*sigma,col=4,lwd=1)
lines(tg,llr$mt-1.96*sigma,col=4,lwd=1)
```