Local linear regression and local Poisson regression

Write a report that contains the results of the computations that you are asked to carry out below, as well as the explanation of what you are doing. The main text (3 or 4 pages) should include pieces of source code and graphical and numerical output.
Upload your answers in a .pdf or .html document (use LaTeX or R Markdown, for instance) to ATENEA, as well as the source code (*.R or *.Rmd, for instance). Your work must be reproducible.

---

1. Estimating the conditional variance by local linear regression

Aircraft Data

We are using *Aircraft data*, from the R library `sm`. These data record six characteristics of aircraft designs which appeared during the twentieth century.

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

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

Go to `R` and charge the library `sm`:

```
library(sm)
```

Now upload the data:

```
data(aircraft)
help(aircraft)
attach(aircraft)
lgPower <- log(Power)
lgSpan <- log(Span)
lgLength <- log(Length)
lgWeight <- log(Weight)
lgSpeed <- log(Speed)
lgRange <- log(Range)
```

1.1. Estimating the conditional variance

Consider the heteroscedastic regression model

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

where $E(\varepsilon) = 0$, $V(\varepsilon) = 1$ and $\sigma^2(x)$ is an unknown function that gives the conditional variance of Y given that the explanatory variable is equal to x.

Let us define $Z = \log((Y - m(x))^2) = \log \epsilon^2$ and $\delta = \log(\varepsilon^2)$. Then

$$Z = \log \sigma^2(x) + \delta,$$

and $\delta = \log \varepsilon^2$ is a random variable with expected value close to 0 (observe that $E(\log \varepsilon^2) \approx \log E(\varepsilon^2) = \log V(\varepsilon) = \log 1 = 0$) taking the role of *noise* in the regression of Z against x (that is, Z is the response variable and x is the predicting variable).

Given that the values of ϵ_i^2 are not observable, a way to estimate the function $\sigma^2(x)$ is as follows:

1. Fit a nonparametric regression to data (x_i, y_i) and save the estimated values $\hat{m}(x_i)$.

2. Transform the estimated residuals $\hat{\epsilon}_i = y_i - \hat{m}(x_i)$:

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

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

$$\hat{\sigma}^2(x) = e^{\hat{q}(x)}.$$

Apply this procedure to estimate the conditional variance of `lgWeigth` (variable Y) given `Yr` (variable x). Draw a graphic of $\hat{\epsilon}_i^2$ against x_i and superimpose the estimated function $\hat{\sigma}^2(x)$. Lastly draw the function $\hat{m}(x)$ and superimpose the bands $\hat{m}(x) \pm 1{,}96\hat{\sigma}(x)$.

Attention: Do the work twice:

- First, use the function `loc.pol.reg` that you can find in ATENEA and choose all the bandwidth values you need by leave-one-out cross-validation (you have not to program it again! Just look for the right function in the `*.Rmd` files you can find in ATENEA)

- Second, use the function `sm.regression` from library `sm` and choose all the bandwidth values you need by *direct plug-in* (use the function `dpill` from the same library `KernSmooth`).

2. Local Poisson regression

2.1. Bandwidth choice for the local Poisson regression

In Atenea you can find the file `03_Band_Choice_Local_Logistic.Rmd`, where the functions `h.cv.sm.binomial` and `loglik.CV` are defined for the local logistic regression. Modify these functions to obtain a bandwidth choice method for the local Poisson regression based on the leave-one-out cross-validation (loo-CV) estimation of the expected likelihood of an independent observation.

Remember that the loo-CV estimation of the expected log-likelihood of an independent observation, when using h as bandwidth, is

$$\ell_{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\widehat{\Pr}_h^{(-i)}(Y = y_i | X = x_i)\right),$$

where $\widehat{\Pr}_h^{(-i)}(Y = y_i|X = x_i)$ is an estimation of

$$\Pr(Y = y_i|X = x_i) = e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!},$$

and

$$\lambda_i = \mathbb{E}(Y|X = x_i)$$

should be estimated by maximum local likelihood using h as bandwidth (for instance, using the function `sm.poisson` from the R package `sm`).

2.2. Local Poisson regression for Country Development Data

Consider the country development dataset (file `HDI.2017.subset.csv` in Atenea) containing information on development indicators measured in 179 countries (Source: Human Development Data (1990-2017), The Human Development Report Office, United Nations, `http://hdr.undp.org/en/data`). Variable `le.fm` always takes non-negative values. Define `le.fm.r` as the rounded value of `le.fm`:

```
le.fm.r <- round(le.fm)
```

Fit a local Poisson regression modeling `le.fm.r` as a function of `Life.expec`. Use `sm.poisson` from the R package `sm` with the bandwidth obtained by loo-CV.