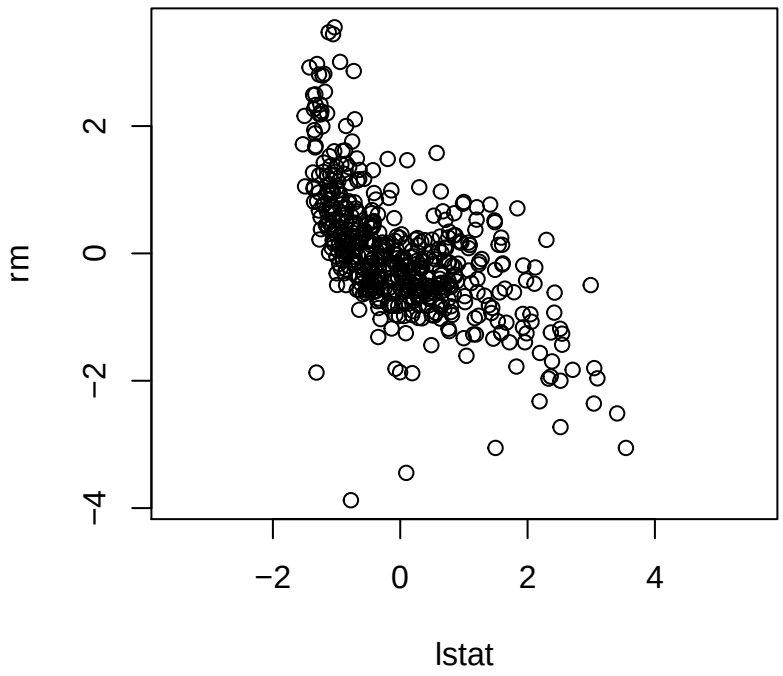
Density estimation. Clustering. (Assignment)

Pedro Delicado

Reading the data: Boston Housing

We'll use the `MASS::Boston` dataset, that contains median house values from Boston neighbourhoods. In particular we are interested in the joint distributions of centered and scaled variables `lstat` and `rm`:

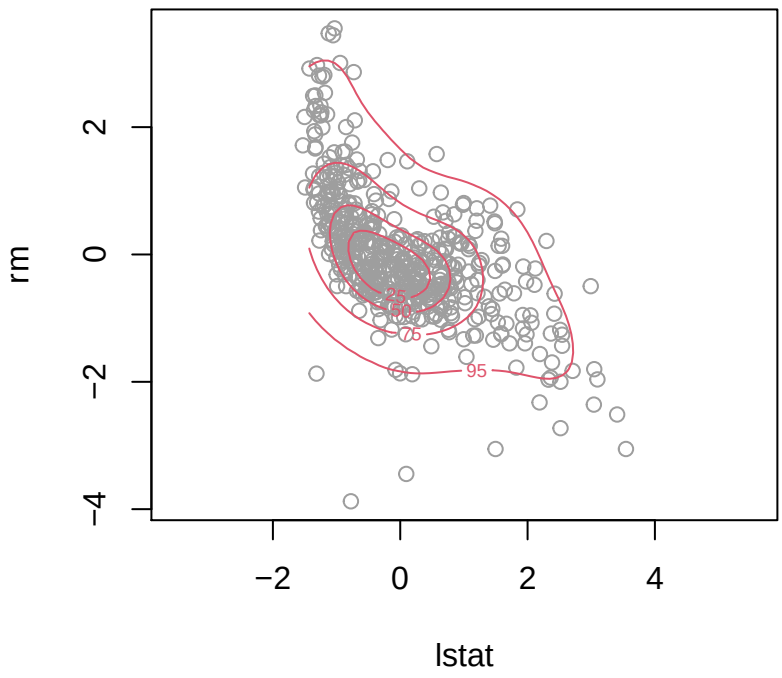
```
data("Boston", package = "MASS")
X <- scale(Boston[,c(13,6)])
plot(X,as=1)
```



Questions

1. We want to estimate the joint bivariate density using a kernel estimator with the same bandwith in both dimensions: $h = (a, a)$. For instance, the following code performs this estimation for $a = 0.5$:

```
library(sm)
plot(X,as=1,col=8)
sm.density(X,h=.5*c(1,1),display="slice",props=c(25,50,75,95),col=2,add=TRUE)
```



Use the *maximum log-likelihood cross-validation method* for choosing the value of a, whem a takes values in the vector `seq(0.1,1,by=0.1)`. Then repeat the previous density estimation using the chosen value of a.

Indication: Maximize in a the **logarithm** of the likelihood cross-validation (instead of maximizing just the likelihood cross-validation). The following code evaluates the logarithm of the density estimator at point $(0, 0)$:

```
new.point <- matrix(c(0,0),ncol=2)
f.hat <- sm.density(X,h=.5*c(1,1),display="none",eval.grid=FALSE,
                    eval.points=new.point)
log(f.hat$estimate)
```

2. Do a hierarchical clustering of these data using the `ward.D` method, plot the resulting dendogram and cut it into $k = 3$ clusters. Plot the scatterplot of the data, using a different color for points in different clusters.

3. For each one of the k clusters obtained above, do the following tasks *(A unique plot should be done, at which the k densities are represented simultaneously)*:

- Consider the bivariate data set of the points in this cluster.
- Estimate non-parametrically the joint density of `lstat` and `rm`, conditional to this cluster *(Use the optimal bandwith found in the first point)*.
- Represent the estimated bivariate density using the level curve that covers the 75% of the points in this cluster.

4. Repeat now points 3 and 4, but choose the number of clusters k according to one (or several) of the automatic criteria we have seen in class. *Optional: If you want, you can choose the optimal bandwidth for each cluster separately (this will improve the final density estimations).*