Challenge B

Rossi Abi-Rafeh

11/15/2017

You will solve, in your group, two challenges for the R programming course. The tasks you have to solve in Challenge B are described in this document. Your submissions for Challenge B will count for the last 25% of your final grade. You have to submit your answers before Friday 8th of December at 11 AM in the morning. In this document, you have all the information about Challenge B.

Rules for Challenge B:

The Challenge is in groups :

- You work with your partners and submit the same answers.
- You have to set up a public Github repository for the project. One Github repo for the group. The github repo has to include all the files needed to compile/knit the final document you submit. We need to be able to fork the github repo, clone it on own computers, run the Rmarkdown file and get the exact same document that you submitted : basically, your final document needs to be perfectly reproducible.
- Submit your final document with the answers on the Moodle challenge link.
- Each member of the group has to submit the answers separately through his/her own Moodle account to receive the grade. Example : Paul and Laura are a team. They make the same pdf file with their answers. Both Paul and Laura submit the document each from their own Moodle account.
- Please detail the time needed to run the steps in Task 3 using the command system.time()

Submissions on Moodle are made of 1 file :

- one pdf/html/ms word file : 2 pages of text (2 pages max - not including the figures/plots) answering the questions, and explaining briefly what you're doing. You have to produce the document using Rmarkdown. It needs to be readable. If you want to comment your code, please do that in the .Rmd file, keep the final document as an answer sheet.
- the document has to include, in the first lines, a link to the github repo of the project

Grading policy :

Steps in this document get you between 1 and 2 points if done correctly. A complicated script that is hard to understand needs to be commented. If it's not, you'll be penalized. Write smart, small, and easy to understand comments.

In the beginning of your R script, do not forget to install and load all the packages you will use later in your script. Test and use the code I gave you for challenge A.

If you load external data, please make the command visible so that we can change the path on our computers when we grade.

As usual, if your .Rmd script does not run on our computers, or shows error messages, your grade is automatically 0. There will be no exceptions to this rule, nor negociations.

If your code runs, and your output does not match what you have in the pdf for a given step, your grade for that specific step will be 0.

For task 3B - it is possible to write simple code that takes max. 10 minutes to run on my personal Mac. If your code takes nights and nights to run, it doesn't qualify as a good answer: please make sure that long code doesn't run automatically when I compile the .Rmd, and instead find a smart way to get some intermediary output inside the .Rmd so that I am able to compile and get the same document as you (think maybe a smaller dataset that you create as an intermediate step or anything else). We will not wait for your .Rmd file to compile for more than an hour, but we still wanna see the code you use to produce the intermediate output.

Task 1B - Predicting house prices in Ames, Iowa (continued)

In this part of Task 1, you will use a common machine learning algorithm to predict the house prices in Ames, Iowa. Non-parametric methods that you study in Intermediate Econometrics are OK and you can choose one of them (implementing a method is a good way to understand how it works...). You will not be graded on the complexity of the method, but rather on whether you can explain nicely what it does, and make a simple implementation of it.

Step 1 - Choose a ML technique : non-parametric kernel estimation, random forests, etc... Give a brief intuition of how it works. (1 points)

Step 2 - Train the chosen technique on the training data. Hint : packages `np` for non-parametric regressions, `randomForest` for random forests. Don't use the variable `Id` as a feature. (2 points)

Step 3 - Make predictions on the test data, and compare them to the predictions of a linear regression of your choice. (2 points)

Task 2B - Overfitting in Machine Learning (continued) - 1 point for each step

Step 1 - Estimate a low-flexibility local linear model on the training data. For that, you can use function `npreg` the package `np`. Choose `ll` for the method (local linear), and a bandwidth of 0.5; Call this model `ll.fit.lowflex`

Step 2 - Estimate a high-flexibility local linear model on the training data. For that, you can use function `npreg` the package `np`. Choose `ll` for the method (local linear), and a bandwidth of 0.01; Call this model `ll.fit.highflex`

Step 3 - Plot the scatterplot of x-y, along with the predictions of `ll.fit.lowflex` and `ll.fit.highflex`, on only the training data. See Figure 1.

Step 4 - Between the two models, which predictions are more variable? Which predictions have the least bias?

Step 5 - Plot the scatterplot of x-y, along with the predictions of `ll.fit.lowflex` and `ll.fit.highflex` now using the test data. Which predictions are more variable? What happened to the bias of the least biased model?

Now let's see what happens to the overall error rate, that is the mean square error. Remember the mean squared error MSE : $MSE^{model} = \frac{1}{n}\sum_i(\hat{y}_i^{model} - y_i)^2$.

Step 6 - Create a vector of bandwidth going from 0.01 to 0.5 with a step of 0.001

Step 7 - Estimate a local linear model y ~ x on the training data with each bandwidth.

Step 8 - Compute for each bandwidth the MSE on the training data.

Step 9 - Compute for each bandwidth the MSE on the test data.

Step 10 - Draw on the same plot how the MSE on training data, and test data, change when the bandwidth increases. Conclude.

Task 3B - Privacy regulation compliance in France

The CNIL (Commission Nationale de l'Informatique et des Libertés) is the French government body that regulates digital freedom, as well as user data protection and privacy. Each company or organization in France wishing to adopt the regulatory framework of the CNIL has to nominate a CIL : a deleguate that will ensure that the internal use of user data is consistent with the CNIL's recommendations, basically a Privacy Officer. This procedure is not mandatory in general, but becomes so if the company wishes to deal with sensitive data (like healthcare.), and is becoming mandatory under the new European Directive for Data and Privacy protection.

DATA : Institutions that nominated CIL

The list of companies and organizations that nominated a CIL is publicly available. However we do not know anything about the size or the sector of these companies.

The SIREN dataset, made public by the french government, compiles a list of all french companies or organizations, as well as details about their employment size, and sector. The original data is available here (Using the 1st november stock datasets)

DATA : SIREN November 1st Stock download link

DATA : SIREN Open Data Portal link

Step 1 - Import the CNIL dataset from the Open Data Portal. (1 point)

Step 2 - Show a (nice) table with the number of organizations that has nominated a CNIL per department. HINT : A department in France is uniquely identified by the first two digits of the postcode. (1 point)

Step 3 - Merge the information from the SIREN dataset into the CNIL data. Explain the method you use. HINT : In the SIREN dataset, there are some rows that refer to the same SIREN number, use the most up to date information about each company. (2 points)

Step 4 - Plot the histogram of the size of the companies that nominated a CIL. Comment. (1 points)

Figures for task 2B

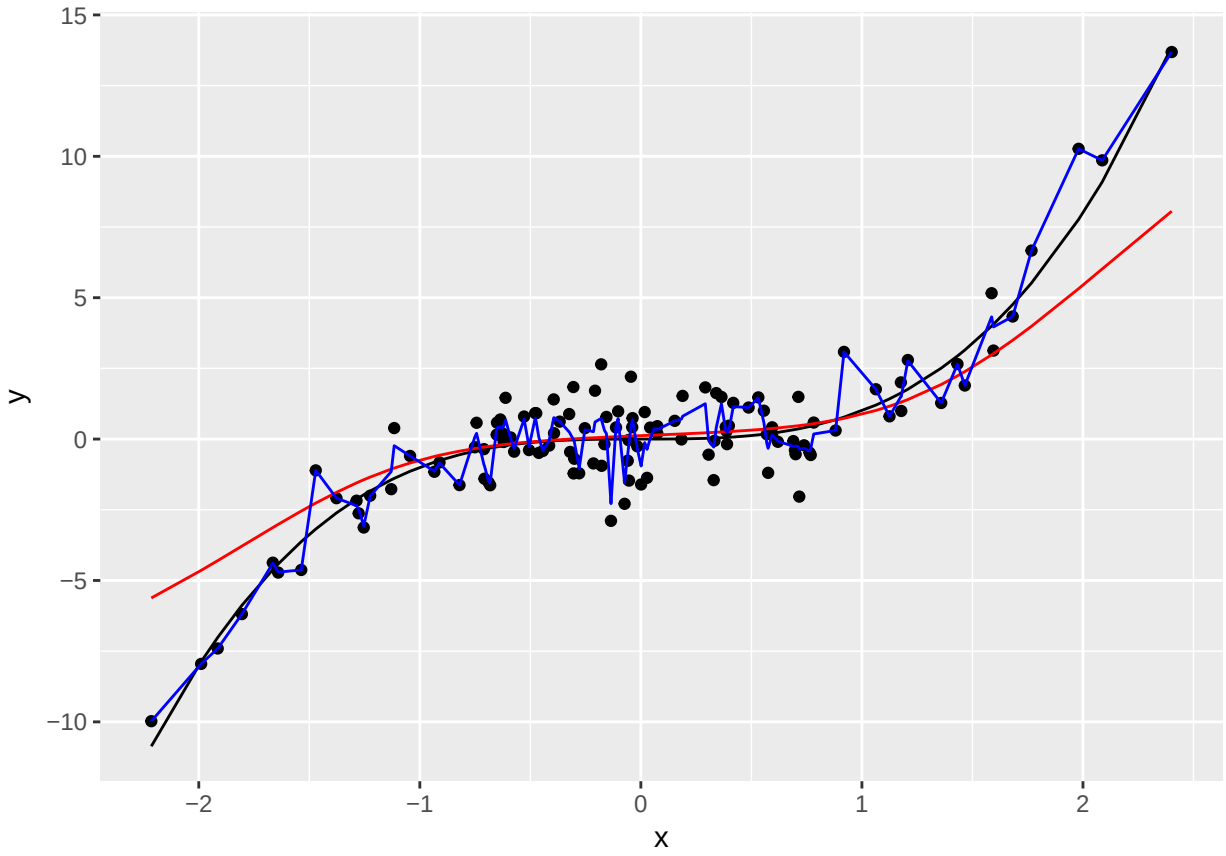

Figure 1: Step 3 - Predictions of ll.fit.lowflex and ll.fit.highflex on training data.

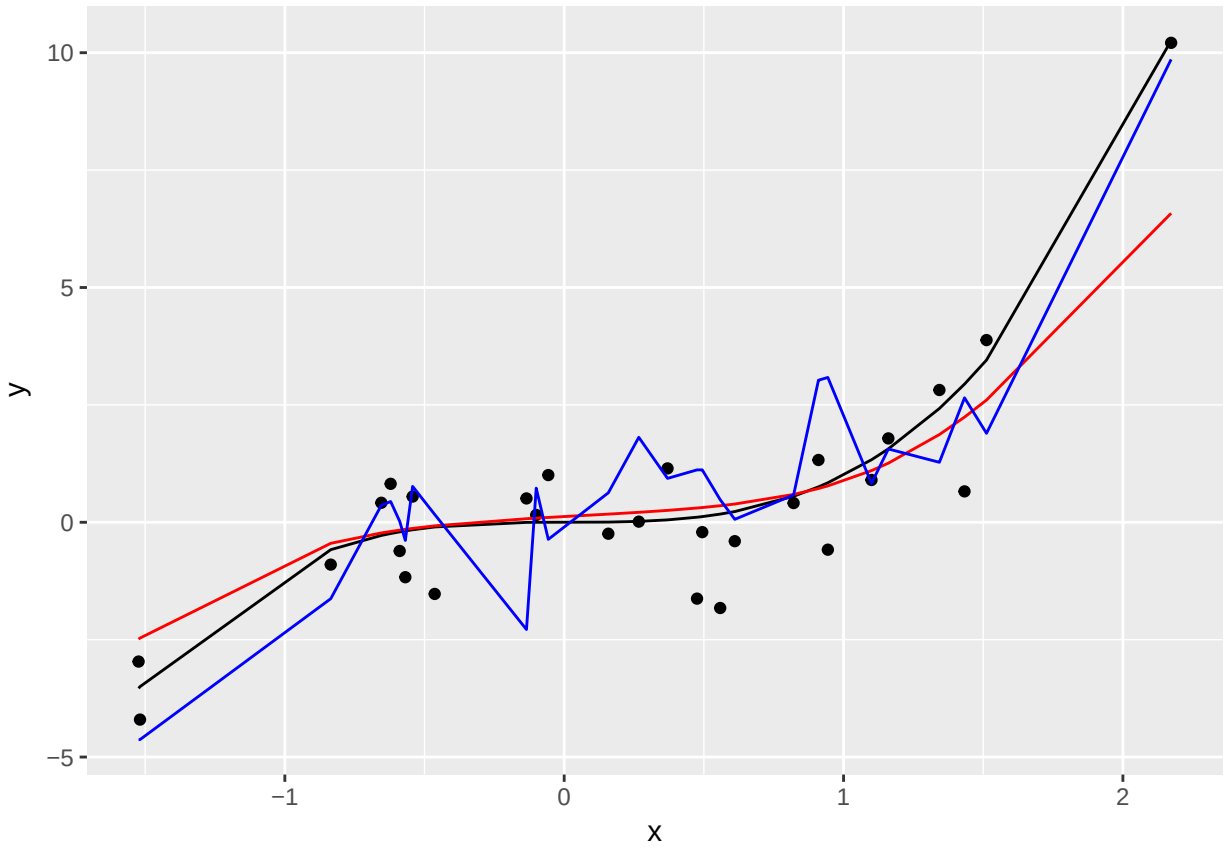

Figure 2: Step 5 - Predictions of ll.fit.lowflex and ll.fit.highflex on test data.

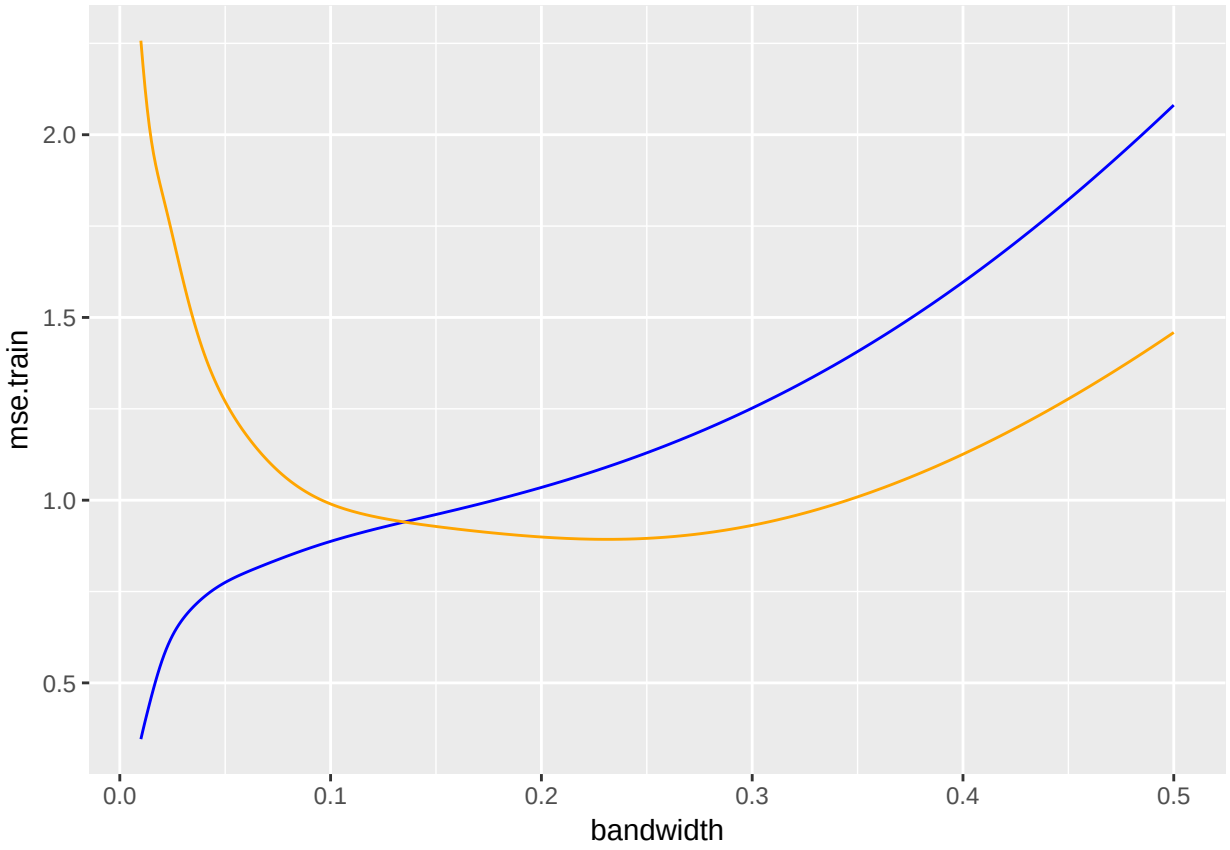

Figure 3: Step 10 - MSE on training and test data for different bandwidth - local linear regression