Présentation MRR

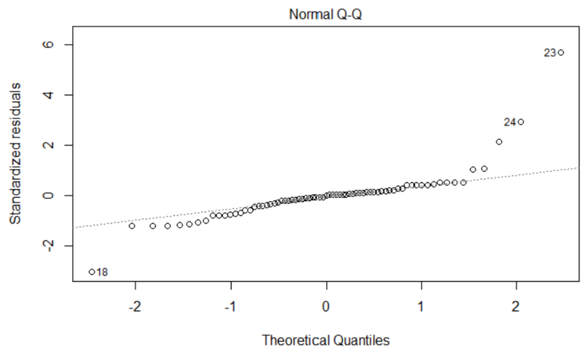
PRUDHOMME Pierre, BERTON Thomas

Projet MRR Ocean
Charge de Projet : Mr.Charentoni

Introduction

After we had done a quick introduction to our variables and try to explain them as much as we could. We are now trying to work on them and to apply few models on them to see which one fit the most. We will be looking into some criteria to try to understand our dataset. Our target variable here is *CHL-5*.

First Linear regression



This was our first try on a linear regression. We have try to modelize a part of the Chlorophylle-A concentration by doing a regular linear regression on the fifth Chlorophylle-A measure. As we can see here, our residuals are following a normal distribution on the normal QQ-plot.

Models

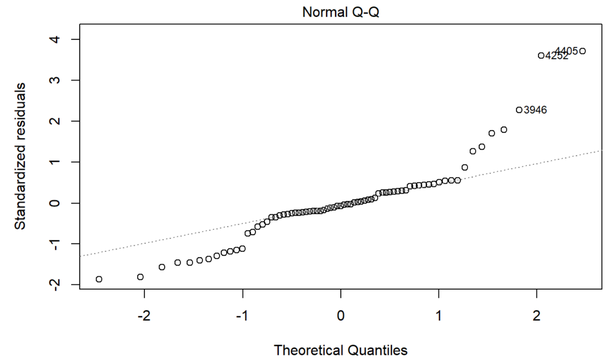
Our objective is to apply few models to our dataset and see the result. So are gonna apply a linear regression but this this time we are going to work on our data before doing it. We normalize our dataset, and then we will apply a ridge and lasso regression.
Once we have work on our data, we need to focus on two other variables. Space and Time, as we know we have to focus on one point BATS(32.10851, -64.01247) and we have 8 others points to study too. And we can also work on the time.

Working on the dataset

We first need to work on our data and to try to make them relevant as much as we can. So we need to create few datasets to work on it with all our variables. And we need to apply a normalization of our data in order to spread out our values. We will also create new dataset who will contains spacial and temporal values.

First Linear Regression Again

After what we have done a regression, we realize that our regression was a bit inappropriate. So we take on one year (2005) and we focus on the BATS points results. We have normalize our data and we use the *'CHL 5'* column and few others to look into it.



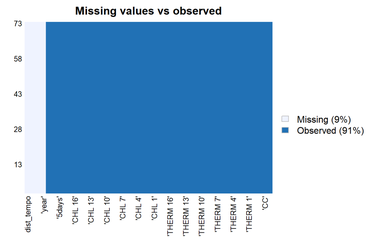
In this model, we do have a $R^2 = 0.8$.
We can assume here that the residuals are maybe following a normal distribution. But let's do a Shapiro-Test to try to get an answer for our hypothesis.

Shapiro-Wilk normality test
W = 0.87923, p-value = 4.35e-06

We do see here that even with a nice plot, we have a p-value around 10^{-6} , so we can't say that our model residuals are following a normal distribution. And now we need to continue to test the next models.

Forward, Backward and Stepwise Selection

After trying unsuccessfully to work with our data with a forward selection, i try to look into missing values. And i figure that few columns contains some 'NA' values on my model.



After applying this selection we could expect to see some of the most variables appeared. So after renaming my columns and working on my dataset. We try to work on our data with a linear model before trying with generalized linear models, and we obtain few results.

With the first generalized linear Model we got :

Generalized Linear Model AIC = 16
Number of Fisher Scoring iterations: 25
Linear Model RMSE = 0.4

The number of Fisher iterations can only allows us to say that our model had merged after few iterations. But we can watch the *Akaike Information Criterion* that is at a low level maybe we can watch all of the selection answer :

Forward / Backward / Stepwise Selection AIC = 6

So we do have 3 times the same AIC here, that doesn't help us a lot. We have few columns that are most significant, they have the smallest p-value or the highest R^2. They end up being *CHL-1* and *5days*. Here we have a lot of degrees of freedom so the treshold is going to be very low.

The residual sum of squares (RSS) is the absolute amount of explained variation, and for the 3 models we got :

Residuals Sum of Squares (RSS) = 73

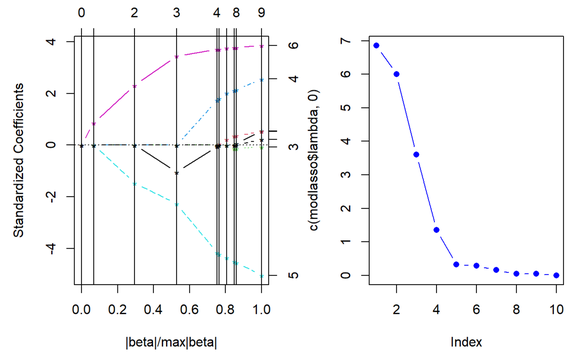
Lasso Method

For the Lasso Method, i focus again on the point BATS and the year 2005 and only left few columns :

SSH, CC, WS, SR, THERM-1, CHL-1, CHL-5, 5days, Cos5days, Sin5days

And after applying the Lasso Method I obtain few results :

Lasso Method AIC = 144.5
Lasso Method RSS = 12.57



So we try to plot our model and to understand our values :

Lasso Method λ = 0.045
Lasso Method RMSE = 0.414

By doing a cross-validation method, this will indicate which variable should we take and this pick the coefficient from the best model, the lambda that minimize the RMSEc. We find the same RMSE as before that doesn't help us a lot in finding which model is better.

Présentation MRR

PRUDHOMME Pierre, BERTON Thomas

Ridge Model

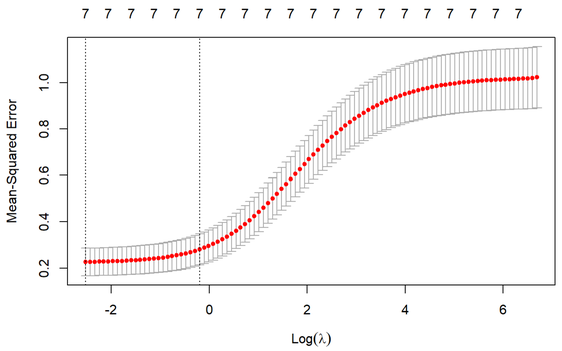
For the Ridge Model, we focus again on the point BATS and the year 2005 and only left few columns :

SSH, CC, WS, SR, THERM-1, CHL-1, CHL-5, 5days, Cos5days, Sin5days

And after applying the Ridge Model we obtain few results :

Ridge Method AIC = 130

Ridge Method RSS = 16.13



We did once again a cross-validation in order to aim at the lowest lambda available :

Ridge Method λ = 0.085

Ridge Method RMSE = 0.414

This time we do have a sighlty bit bigger lambda but we find a RMSE that is close to our previous results.
That doesn't help us a lot, that means our parameters (lambda) is almost the same in Lasso and Ridge Method, and the RSME too we cant fix the choice of our model by doing this.

Results for Classic DataSet

We have a part of our dataset that is constitued of :

SSH, CC, WS, SR, THERM-1, CHL-1, CHL-5, 5days, Cos5days, Sin5days

Our target variable was the *CHL-5* and we choose to keep few columns that contains the value of the measurement of climatic conditions. After scaling all of this data, we apply 3 different model on this dataset : *Lasso*, *Ridge*, and a classic Linear Model.

By doing this we got few results, the *RMSE* (Root-Mean-Square Error) that is almost equal for every model. Our goal is to minimize it and in our work we got always a RMSE close to 10^{-1} so we can't choose our model on this criteria.
The *AIC* (Akaike Information Criterion) was also close in every model, about 10^1 our goal is also to minimize it, with this type of variation in AIC we can't say which model is better.
And by doing *Lasso* and *Ridge* method we aim to find a lambda that maximize the model, and we also have a similar lambda in both models.

Spatial Model

For this dataset, we added all the *CHL-1*, *SSH*, *CC*, *THERM1*, *WS*, *SR*, *5days-cos* and *5days-sin* columns to our dataset. So now we have all the informations of the 9 points and we have the *CHL-5* for the point BATS

Ridge Model

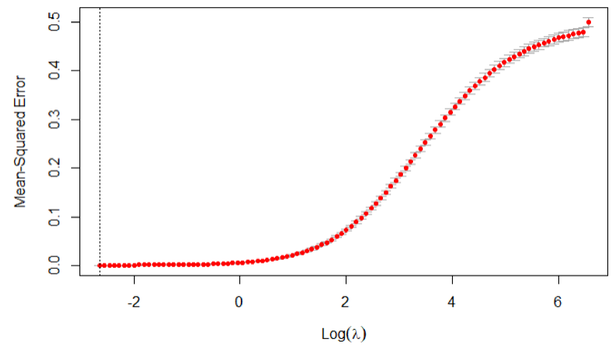
For the Lasso Method, we focus again on the information of all points and only left few columns :

SSH, CC, WS, SR, THERM-1, CHL-1, CHL-5, 5days, Cos5days, Sin5days

And after applying the Ridge Model we obtain few results :

Ridge Method AIC = 22.6053

Ridge Method RSS = 66.68472



We did once again a cross-validation in order to aim at the lowest lambda available :

Ridge Method λ = 0.03

Ridge Method RMSE = 0.0924842

This time we do have a sighlty bit bigger lambda but we find a RMSE that is close to our previous results.
That doesn't help us a lot, that means our parameters (lambda) is almost the same in Lasso and Ridge Method, and the RSME too we cant fix the choice of our model by doing this. However, the lambda is smaller in lasso so lasso is maybe a bit better.

Lasso Method

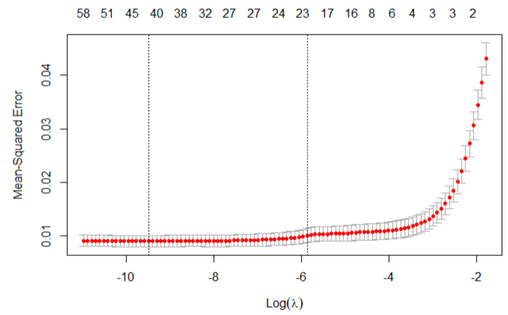
For the Lasso Method, we focus again on the information of all points and only left few columns :

SSH, CC, WS, SR, THERM-1, CHL-1, CHL-5, 5days, Cos5days, Sin5days

And after applying the Lasso Method we obtain few results :

Lasso Method AIC = 162.4542

Lasso Method RSS = 9.818199



So we try to plot our model and to understand our values :

Lasso Method λ = $9.923074e-05$

Lasso Method RMSE = 0.08894674

By doing a cross-validation method, this will indicate which variable should we take and this pick the coefficient from the best model, the lambda that minimize the RMSEc. We find an RMSE lower than before. Therefore, the model is more accurate than before.

Results for Spatial Model

We have a part of our dataset that is constitued of :

SSH, CC, WS, SR, THERM-1, CHL-1, CHL-5, 5days, Cos5days, Sin5days.

The CHL-5 is the one from BATS but all the others variables are the ones from the 9 points.

Our target variable was the *CHL-5* and we choose to keep few columns that contains the value of the measurement of climatic conditions. After scaling all of this data, we apply 2 different models on this dataset : *Lasso* and *Ridge*.

By doing this we got few results, the *RMSE* (Root-Mean-Square Error) which is almost equal for every model. Our goal is to minimize it and in our work we got always a RMSE close to 10^{-2} so our new dataset gives us better informations than the classic one.
However, the *AIC* (Akaike Information Criterion) was very different in every model, about 10^2 in lasso and 10^1 in Ridge. Our goal is also to minimize it. Therefore, with this type of variation in AIC we may prefer Ridge over lasso.

Temporal Method

For this dataset, we added all the *CHL-1*, *SSH*, *CC*, *THERM1*, *WS*, *SR*, *5days-cos* and *5days-sin* columns +/-k5days to our dataset. Unfortunately, we had a probleme with the construction of the dataset and therefore we weren't able to get decent graph and values concerning this method. Our job is now to solve this issue in order to study this model.