Achddou Mathis Skema Business School

Final project: data analysis in R programing

For this work we wanted to stay in the market finance area. More specifically working on fixed income securities. To this extent, we chose to focus on the “OAT 10 years”, a bond emitted by the French government to finance its projects.
This security is quite important from a global macroeconomic point of vue as it represents the second most exchanged and bought bond from the 27 of the European Union behind the “Bund 10 years” originated by the German government.

Our statistical question is straight forward:

Between the Lasso, Probit and OLS models, which one do better forecast the yield of the OAT 10 Years?

Table des matières

Data selection and cleaning

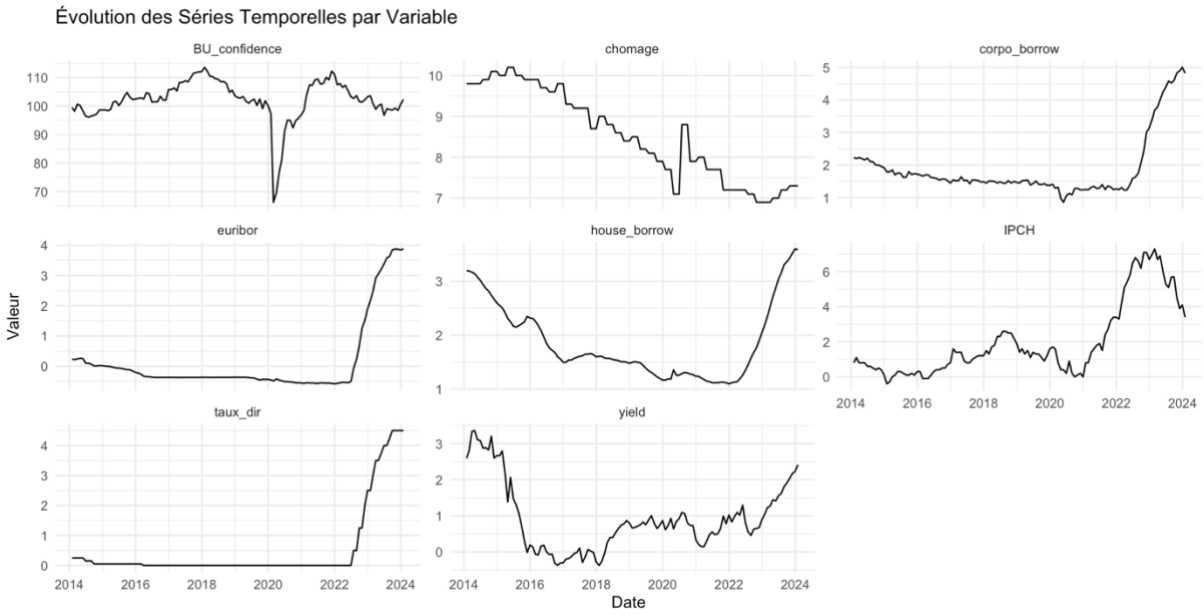
Most of the work to be done is on the selection of the data and cleaning this latter. We tried to find representative and relevant data to apply our model.
We chose 10 years of data from January 2014 to January 2024, selecting one data point per month (example: each last day of trading of each month).

The first step was to find adequate data for the OAT 10y for the indicated dates. For the explicative variables we chose

- Cost of borrowing for corporation
- Inflation in France between 2014 and 2024
- The unemployment rate in France
- The cost of borrowing for households
- The ceiling rate of the European Central Bank
- Business confidence
- Forward EURIBOR 1 month

As each of these data are from different sources, they were all differently presented (daily, weekly, quarterly, …) which was a real struggle to normalize. We first wanted to apply our model on a daily basis as yields are quoted from openings to closings of markets on every trading day.
The monthly approach was selected thanks to the advice of our professor, this approach helped us greatly simplify the work as most of our data were not daily.

However, many variables were rejected as they were published yearly and not monthly such as the Gross Domestic Product.



Above are the plots of our variables. Visualizing helps us understand better the data and even preview the analytics we will find after.

Descriptive statistics

First, we will check the mean, median, mode, standard deviation, skewness, and kurtosis of the data selected:

	Yield	Euribor_1m	Corpo_borrowings	IPCH
Mean	0.9175785124	0.1318181818	1.879504132	1.990082645
Median	0.735	-0.369	1.53	1.3
Mode	2.669	-0.372	1.52	0.8
Standard deviation	0.9462224347	1.199377165	0.9425805458	2.117443622
Skewness	0.9567268481	2.277072691	2.148280699	1.241380627
Kurtosis	3.157233504	6.769500873	6.55905669	3.310050632

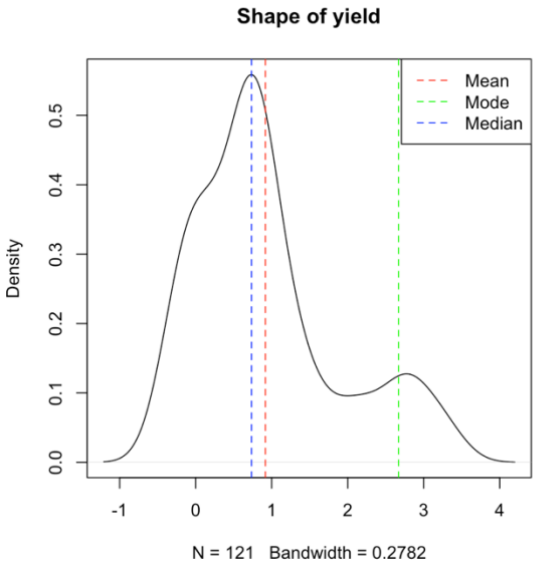
	Unemployment	House_borrowings	Key_interest_rate	Business_confidence
Mean	8.527272727	1.86677686	0.5090909091	102.1553719
Median	8.5	1.59	0	102.4
Mode	7.2	1.5	0	100.2
Standard deviation	1.103856875	0.6965129993	1.251881917	7.209576391
Skewness	0.01318354191	0.9285842331	2.414823141	-2.159751334
Kurtosis	1.54440159	2.700859677	7.229098689	11.09386667

Most of the results are not surprising as post 2014 was a long period of low interest rates resulting from the economic crisis. These results were even forecastable if the macroeconomic and rate policies from 2014 to 2024 are known by the reader.

Analysis of the results

1. Yield

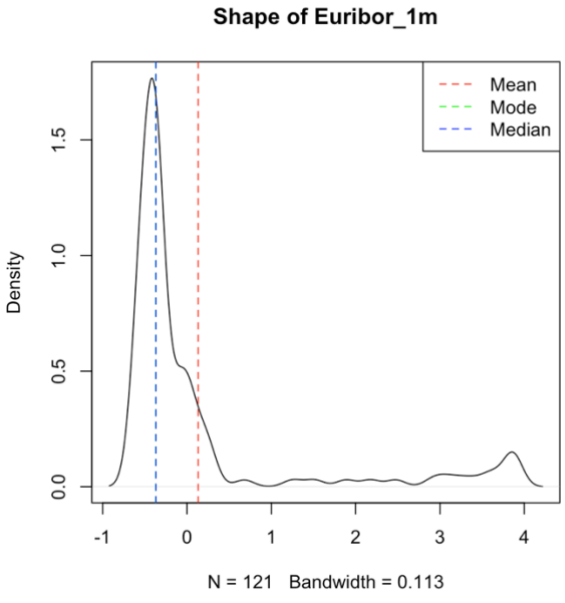
A moderately positive skewness in yield could indicate occasional periods of unusually high returns. The kurtosis close to normal suggests that extremes are not too frequent, indicating a relatively stable market with occasional peaks in returns that savvy investors might exploit.



Here is representation of the density curve of the "Yield". As the mean is superior to the median the distribution of the data presents a positive skewness and asymmetry.

2. Euribor 1 month

The high kurtosis and significant positive skewness indicate a sharp distribution with extremely low or negative rates more frequent than expected. This may reflect periods of very accommodative monetary policy by the ECB.



Here is representation of the density curve of the "Euribor_1m".
The Mode and the Median are almost the same, hence we see no representation for the Mode.
As the mean is superior to the median the distribution of the data presents a positive skewness and asymmetry.

3. Corporations Borrowing costs

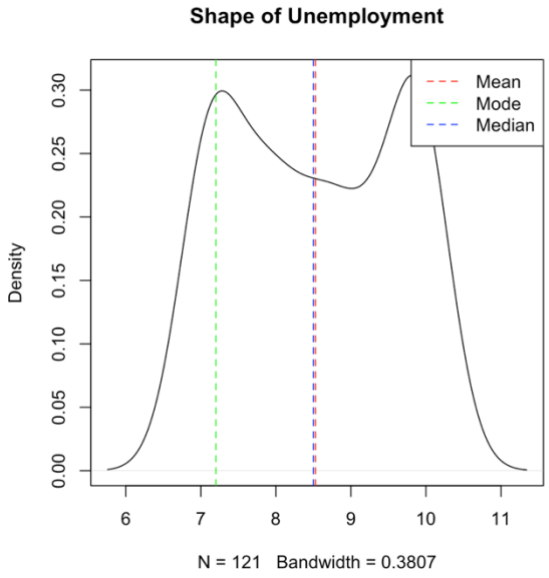
The high skewness and kurtosis suggest that borrowing costs for corporations can sometimes be much higher than normal, perhaps reflecting credit crises or sudden changes in risk perception.

4. IPCH

Moderate skewness and slightly elevated kurtosis might reflect episodes of sudden inflation, though the distribution remains relatively normal.

5. Unemployment

Skewness and kurtosis values close to normal, indicate a symmetric and regular distribution of the unemployment rate, without extreme variations.



Here is representation of the density curve of the “Unemployment”.
The Mean and the Median are almost the same.
An almost perfect symmetry is seen here, it is linked to the fact that the Mode is below the Mean and Median which are equal.

6. House Borrow

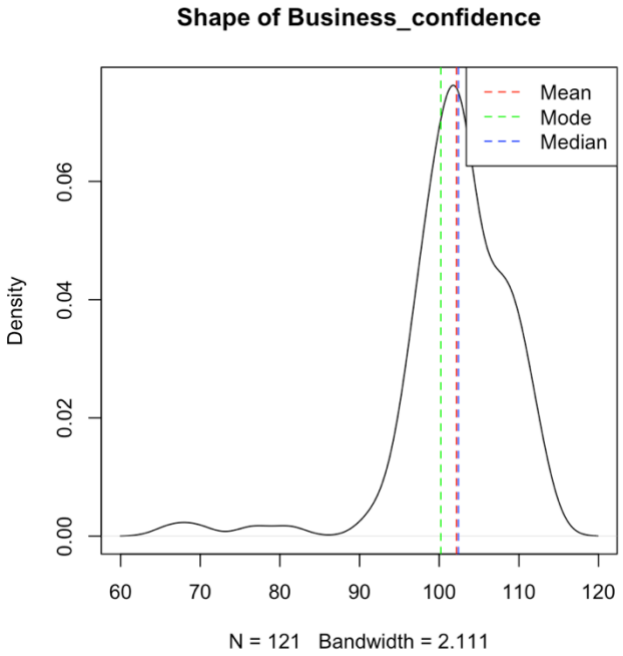
Slight skewness and kurtosis below three indicate that rates for housing loans are generally stable with few extreme events.

7. ECB Policy Rates

The extreme kurtosis and positive skewness indicate periods when ECB policy rates have remained at low or zero levels, with occasional spikes. This may indicate periods of monetary policy aimed at stimulating the economy through very low interest rates, useful for analysts and economists forecasting economic trends.

8. Business Confidence

The very high kurtosis and negative skewness indicate a distribution with extreme values on the lower side more frequently than usual, perhaps due to economic or political crises negatively affecting confidence.



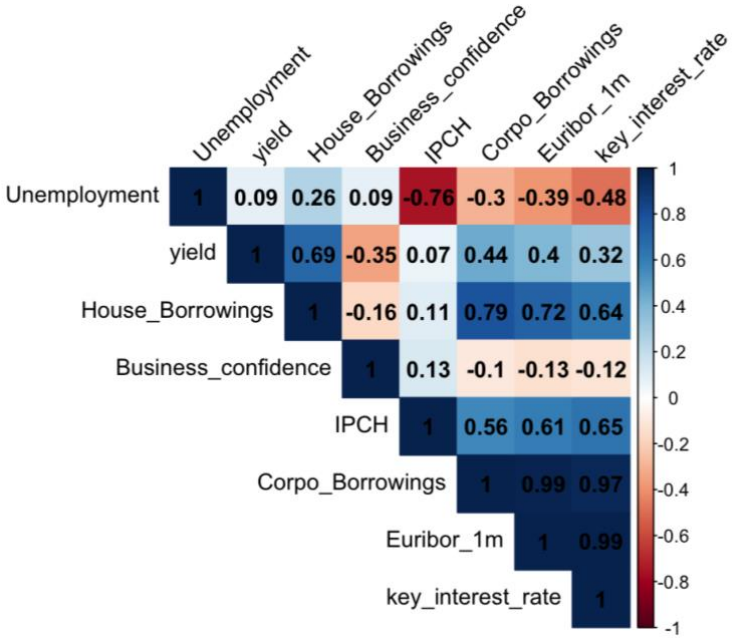
Here is representation of the density curve of the “Business Confidence”.
The Mean and the Median are almost the same.
The mode, the mean and the median being very near, and having a long tail on the left side, we can conclude that the skewness is negative.

Conclusion: The high skewness and kurtoses in some variables suggest that standard models may sometimes be insufficient to capture underlying dynamics, and more nuanced analyses may be necessary for a comprehensive understanding.

Are the necessary assumptions of an Ordinary Less Squared regression respected?

1) Correlation between variables

The **first** one to be tested is the **correlation** between the explicative variables. We will **then** check the **multicollinearity** to make sure both results point to the same direction. To this extent we will analyze both correlation matrix and VIF scores.



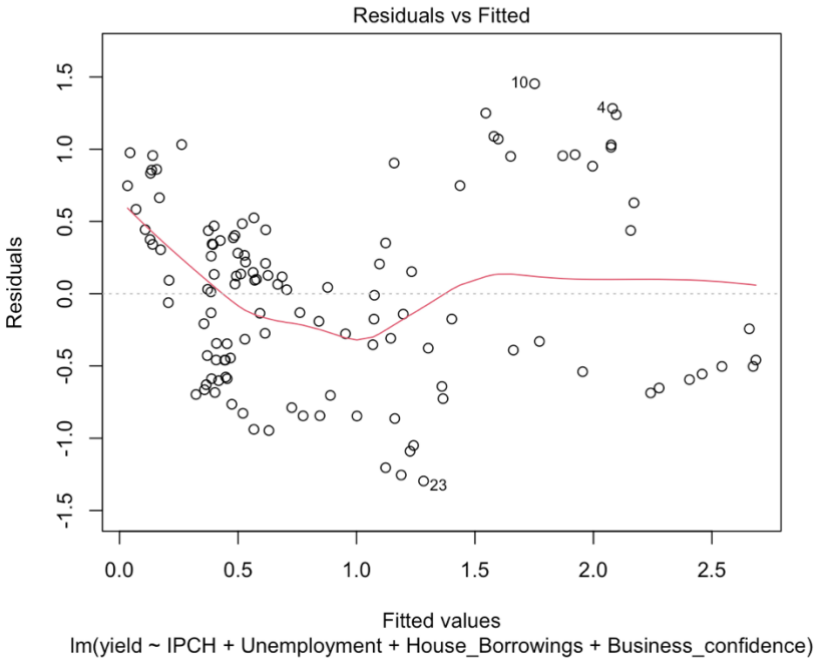
Hopefully we checked the correlation before going deep inside the subject. We clearly see here that 3 variables are strongly correlated, thus they are useless and harming for the analysis.

We decided to delete “Euribor 1m”, “Corporate borrowings” and “….”.

2) Checking the assumptions with the new variables

We will now check the OLS assumptions on our model without the 3 variables identified above.

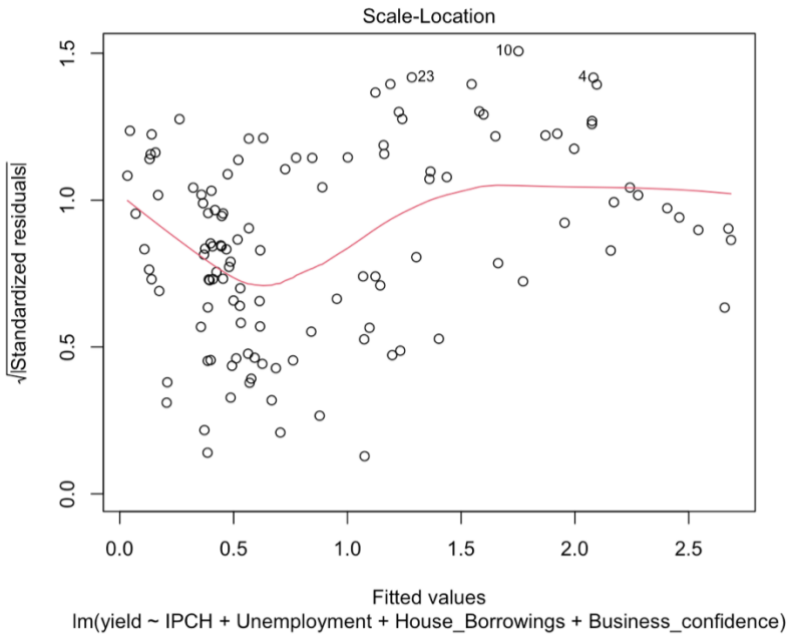
- **Linearity**, the relationship between the independent variables and the dependent variable must be linear. This means that changes in the explanatory variables result in proportional changes in the response variable. You can verify this assumption by using scatter plots to observe the relationships or by examining the residuals from the regression.



The linearity is not perfect but is far from being ridiculous as well.

We consider the linearity assumption as validated.

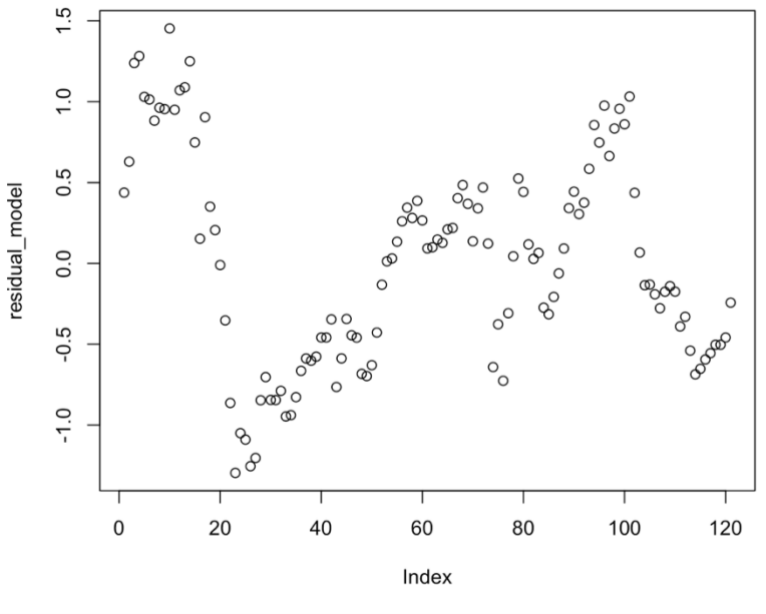
- **Homoscedasticity**, the variance of the residuals (errors) should be constant across all predicted values. This means that the residuals should have consistent dispersion across the regression line. Non-constant variance (heteroscedasticity) can be detected through residual plots.



Residuals are almost spread equally along the ranges of predictors which is a condition for validating the homoscedasticity.

We consider the homoscedasticity assumption as validated.

- **<u>Independence of errors</u>**, the residuals should be independent of each other, which means there should be no correlation between consecutive residuals. This assumption is crucial to avoid autocorrelation, especially in time-series or longitudinal data. Tests such as the Durbin-Watson test can be used to detect autocorrelation.



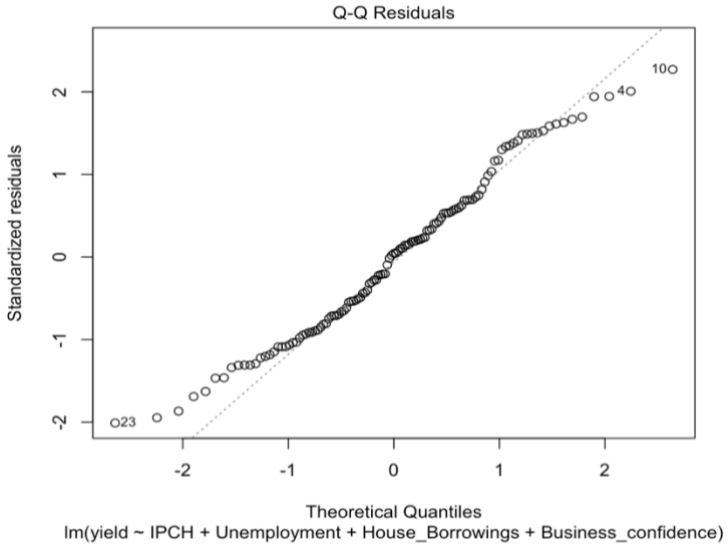

```
        Durbin-Watson test

data:  reg1
DW = 0.14535, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The plot is suggesting an autocorrelation. The Durbin Watson test will help us confirm or not this suggestion.

The test confirmed our intuition, the assumption is not validated.

- **<u>Normality of residuals</u>**, for optimal validity of statistical tests (like tests on coefficients), the residuals of the regression should ideally follow a normal distribution. While linear regression can technically be applied without this condition, for sufficiently large samples (under the central limit theorem), it helps ensure the reliability of confidence interval estimates and p-values. Tests like the Shapiro-Wilk test or histograms of residuals can help assess this normality.



The points fall approximately around the line, we can assume that the normality of residuals is respected.

The assumption is validated.

- **No Multicollinearity**, the explanatory variables should not be too highly correlated with each other (multicollinearity), as this can make the regression coefficients' estimates unstable and difficult to interpret. You can check for multicollinearity by examining the Variance Inflation Factor (VIF) for each predictor; a VIF greater than 5 typically indicates a problem.

```
> # 1. MULTICOLINEARITY with VIF
> vif(reg1)
                IPCH         Unemployment     House_Borrowings Business_confidence
            3.904048             4.109548             1.671213             1.332173
```

The VIF test is under 5 so we can consider this assumption as validated as well.

Conclusion:

Amongst the 5 assumptions the only one not validated is independence of errors.

What are the results of an OLS applied to our data?

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.031313   1.011012   2.998  0.00332 **
IPCH               -0.039918   0.055543  -0.719  0.47379
Unemployment       -0.122778   0.109312  -1.123  0.26368
House_Borrowings    0.957432   0.110477   8.666  3.1e-14 ***
Business_confidence -0.027161   0.009529  -2.850  0.00517 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.652 on 116 degrees of freedom
Multiple R-squared:  0.541,     Adjusted R-squared:  0.5251
F-statistic: 34.18 on 4 and 116 DF,  p-value: < 2.2e-16
```

These are the results we get after applying an OLS to our data. With an R-2 of 0.541 and an adjusted R-2 of 0.5251.

Before getting rid of the 3 variables, we had an R-2 superior of 0.7, showing that not respecting the assumptions results in artificially boosting the R-2.

Let’s break down the analysis:

Coefficients:
The Intercept is estimated at 3.0313, meaning when all the independent variables are 0, the expected value of the dependent variable is 3.0313. IPCH has an estimated coefficient of -0.039918, indicating a slight decrease in the dependent variable for each unit increase in this variable. Unemployment has an estimated coefficient of -0.122778, also suggesting an inverse relationship with the dependent variable. House_Borrowings has a coefficient of 0.957432, showing a significant increase in the dependent variable for each unit increase in House_Borrowings. Business_confidence has a coefficient of -0.027161, indicating a slight inverse relationship with the dependent variable.

Standard Error:
The standard errors indicate the average variability of the coefficient estimates; the lower they are, the more confident we can be in our coefficient estimates. T values and P values (Pr(>|t|)): The t-value indicates how many standard errors the coefficient estimate is away from 0. The higher the value (in absolute terms), the more significant the result is considered. The associated p-values tell us the probability of observing our data if the true effect of the variable was zero. P-values below 0.05 are typically considered

statistically significant. In our model, House_Borrowings and the Intercept are statistically significant ($p < 0.05$), as is Business_confidence ($p < 0.01$).

Significance Codes:
They indicate the level of significance of the coefficients, with '' indicating $p < 0.001$, '' for $p < 0.01$, and '' for $p < 0.05$.
The R-squared of 0.541 means that the model explains 54.1% of the variance in the dependent variable. The Adjusted R-squared is 0.5251, which is an adjusted version of R-squared that accounts for the number of predictors in the model. This also indicates that the model explains about 52.51% of the variance. The F-statistic is used to test the hypothesis that all regression coefficients are equal to zero (i.e., no effect). A high F-statistic and a very low associated p-value (< 2.2e-16) reject the null hypothesis, suggesting that the overall model is statistically significant.

In summary, our regression model appears to be significant with some predictors having a more notable impact on the dependent variable. House_Borrowings appears to be the most influential predictor with the highest statistical significance in this model.

LASSO model:

This is a penalized regressions that do not require the conditions of linearity and non-multicollinearity of linear regression. Indeed, these techniques add a penalty term to the cost function of linear regression, which improves prediction when the assumptions of linear regression are not met.
The LASSO regression incorporates a penalization of the coefficients by forcing some of them towards 0, and thus it implements a variable selection. The penalty implemented is proportional to the parameter λ, which is the hyperparameter of the model. An optimal selection of λ is crucial to ensure good model performance. This is generally done by cross-validation.

Below are the hypotheses to verify:

Linearity:
As with ordinary linear regression, LASSO assumes that the relationship between the independent variables and the dependent variable is linear. It is therefore important to ensure that the necessary transformations are applied to meet this hypothesis if the relationship is not linear.

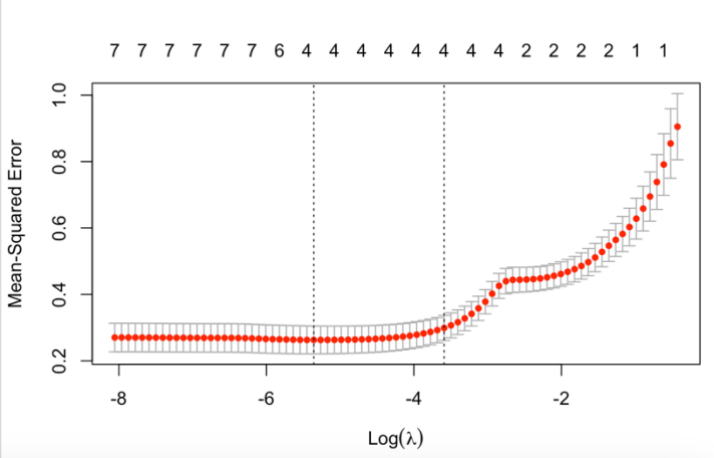
Centering and scaling predictors:
It is crucial to standardize (center and reduce) the explanatory variables before applying LASSO, as the L1 penalty depends on the scale of the variables. If the variables are not on a comparable scale, the penalty will not be applied uniformly, which could distort the selection of variables.

Independence of observations:
Observations must be independent of one another for the estimates to be valid. Dependence between observations can lead to biased estimates.

Absence of perfect multicollinearity:
Although LASSO can manage a certain form of multicollinearity (through variable selection), the presence of perfect multicollinearity (i.e., when variables are exactly linearly dependent) can still pose a problem. It is important to check for and manage this, perhaps by removing or combining highly correlated variables.

Correct model specification:
As with any statistical model, it is essential that the model is correctly specified, meaning that all relevant variables are included and non-relevant variables are excluded. LASSO can help identify non-relevant variables, but it is always wise to think carefully about model specification before analysis.

Below is the result for the lasso model:



The results of crossed validation to find the optimal lambda.

```
(Intercept)          0.9175785
Euribor_1m           .
Corpo_Borrowings     .
IPCH                 .
Unemployment        -0.7583319
House_Borrowings     1.4599640
key_interest_rate   -1.0016847
Business_confidence -0.1375339
> r_squared<-1-sum((predictions-y)^2)/sum((y-mean(y))^2)
> r_squared
[1] 0.7255561
```

The significant predictors are Unemployment, House_Borrowings, and key_interest_rate, with Unemployment negatively associated and the others positively associated with the dependent variable. The R-squared value of 0.725561 suggests that approximately 72.56% of the variability in the dependent variable is explained by the model, which is a relatively strong fit. The absence of coefficients for some variables (indicated by dots) suggests that the LASSO procedure has effectively shrunk their coefficients to zero, deeming them not useful for the model.

Probit model

The assumptions of linear regression are not all met, which implies that the results may be unreliable. Therefore, we will try to modify our problem statement to find a model that fits our data better. In this model we standardized our data.

For this, there is Probit regressions, which do not require as many assumptions as linear regression and are thus more flexible. However, these regressions predict a probability, rather than a continuous value. Therefore, we must transform our target variable 'yield' into a binary variable to implement these models. Transforming the dependent variable to represent a probability means modifying our problem statement. Instead of trying to predict the future borrowing rate, we could attempt to determine the probability that the borrowing rate exceeds 1.5% based on different financial indicators, for example.

This issue seems to be equally interesting since a low borrowing rate is ideal for borrowers, and when it increases too much, it can have a lot of impacts. We could be interested in knowing the probability that 'yield' exceeds 1.5% since, when we observe our data, we can notice that 'yield' was higher than 1.5% in the years 2014-2016 and then experienced a significant decrease, becoming negative. But since the end of 2022, it seems to undergo a sharp increase that does not seem to stop, and it could be interesting to try to predict if this trend will continue over time or will experience a new decline, taking into account financial factors that directly influence the borrowing rate. We have also been interested in knowing the probability that the variation of 'yield' exceeds x% since, when we observe our data, we can notice that 'yield' undergoes different variations (positive or negative) at different scales.

In this model we also have hypothesis to verify, which have been done in the code.

- Binary Dependent Variable
- Linearity in Latent Variable
- Normal Distribution of Errors
- Independent Observations
- No Perfect Multicollinearity
- Correct Specification of the Model
- Homoscedasticity

Below is the result for the Probit test:

```
        1          2          3          4          5          6          7          8          9
0.36165210 0.38872569 0.33741332 0.34741790 0.38661612 0.42516745 0.44989782 0.44111678 0.43455478
       10         11         12         13         14         15         16         17         18
0.45566772 0.42501690 0.42719008 0.40922346 0.41568414 0.40707672 0.38291553 0.37663586 0.40840826
       19         20         21         22         23         24         25         26         27
0.35068994 0.31136520 0.28714854 0.30509286 0.32037101 0.31651350 0.31268107 0.30887408 0.31651350
       28         29         30         31         32         33         34         35         36
0.25062515 0.25736037 0.30959458 0.29403497 0.29403497 0.25853369 0.31018417 0.31018417 0.24451720
       37         38         39         40         41         42         43         44         45
0.18955639 0.18150022 0.19647690 0.14930901 0.14930901 0.14266226 0.14706552 0.13111829 0.11666380
       46         47         48         49         50         51         52         53         54
0.08502947 0.08365684 0.08297811 0.08778679 0.09748966 0.11249845 0.10177769 0.10760720 0.10932591
       55         56         57         58         59         60         61         62         63
0.10455634 0.10372724 0.10965478 0.12469410 0.11713920 0.13680691 0.15138465 0.15368216 0.14687435
       64         65         66         67         68         69         70         71         72
0.14039243 0.14919361 0.13932337 0.12671515 0.14997339 0.12573500 0.14591778 0.11755237 0.13518721
       73         74         75         76         77         78         79         80         81
0.14857298 0.73457664 0.67364173 0.42869819 0.33270880 0.16888794 0.30431624 0.30620092 0.35735031
       82         83         84         85         86         87         88         89         90
0.19881182 0.18637806 0.17074441 0.16128494 0.10719672 0.08020045 0.06750676 0.05762298 0.05619220
       91         92         93         94         95         96         97         98         99
0.06530318 0.06369459 0.05343035 0.04231781 0.03245838 0.03567271 0.04888187 0.04725544 0.05273746
      100        101        102        103        104        105        106        107        108
0.04971487 0.05881618 0.06946190 0.06871304 0.06323960 0.07523847 0.06779722 0.06343971 0.05787355
      109        110        111        112        113        114        115        116        117
0.05643693 0.07066000 0.08325980 0.08013343 0.07628998 0.10458661 0.09781536 0.10019344 0.10180724
      118        119        120        121
0.10205962 0.10876177 0.09194826 0.08008803
```

According to this model there are more probabilities that the yield is going to be lower than 1.5%

General conclusion:

After trying 3 different models (excluding Probit because it answers another issue) for our data set we can conclude that the best fit is the Lasso one with a R-2 of 0.72, 0.2 points more than the linear regression which is not adapted here. Moreover, the hypothesis of Lasso are respected contrary to the OLS model.

Lasso's score indicates a better ability to explain the variance of the dependent variable. This suggests that Lasso, with its coefficient penalization technique to reduce overfitting and eliminate uninformative variables, is more effective for our dataset. To make accurate prediction this model must be preferred compared to OLS.