Case study

Name: Maciej Sliz-Kondratowicz

Task:

*„…let's try to find a model that **uses US treasury rates to describe the deposit rate paid**. However, rather than a full essay, please just set up a plan (key steps) …"*

Solution:

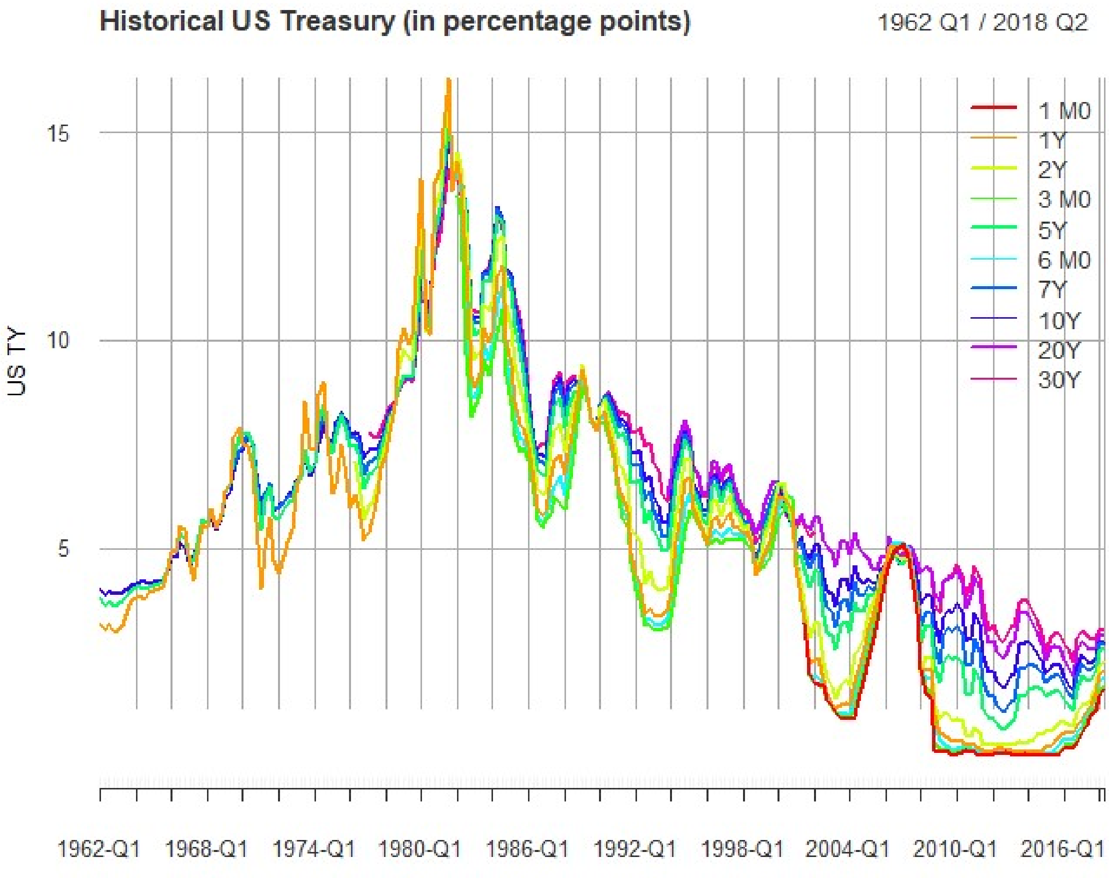
Q1 Warm ups:

***a):** As you may notice, the data come with different frequency. Please write code to clean the data and convert them to the desired frequency.*

Indeed, the information on deposits and interest expense are stored in quarterly frequency, whereas US Treasury Yields are saved in daily frequency. In the first part of *solution.R* script, the historical X variables are transformed from daily to quarterly frequency. Within this task, the NA rows are excluded and quarterly US Treasury Yields are calculated as quarterly average.

As alternative, quarterly end aggregation was considered. However, given that interest is paid through the whole quarter, quarterly average aggregation should better reflect the dynamics of deposit rate paid.

Figure 1 Historical quarterly US Treasury Yields



***b):** “Deposit Rate Paid” is not a provided variable. To get you started, relevant data fields to calculate deposit rate paid by all FDIC insured banks have been downloaded into the attached data file “Deposit Interest.xlsx”. The column “Total interest expense” is the total interest paid in dollar amount. There are a few related balance columns. Please choose the most appropriate balance to use and discuss briefly why you think it is the appropriate one. For the remaining questions, you can just use the “Deposit Rate Paid” variable as you defined here.*

The quarterly Deposit Rate Paid (dependent variable) is calculated as:

Deposit Rate(t) = Total Interest Expense(t) / (Domestic deposits interest-bearing(t) + Foreign deposits(t))

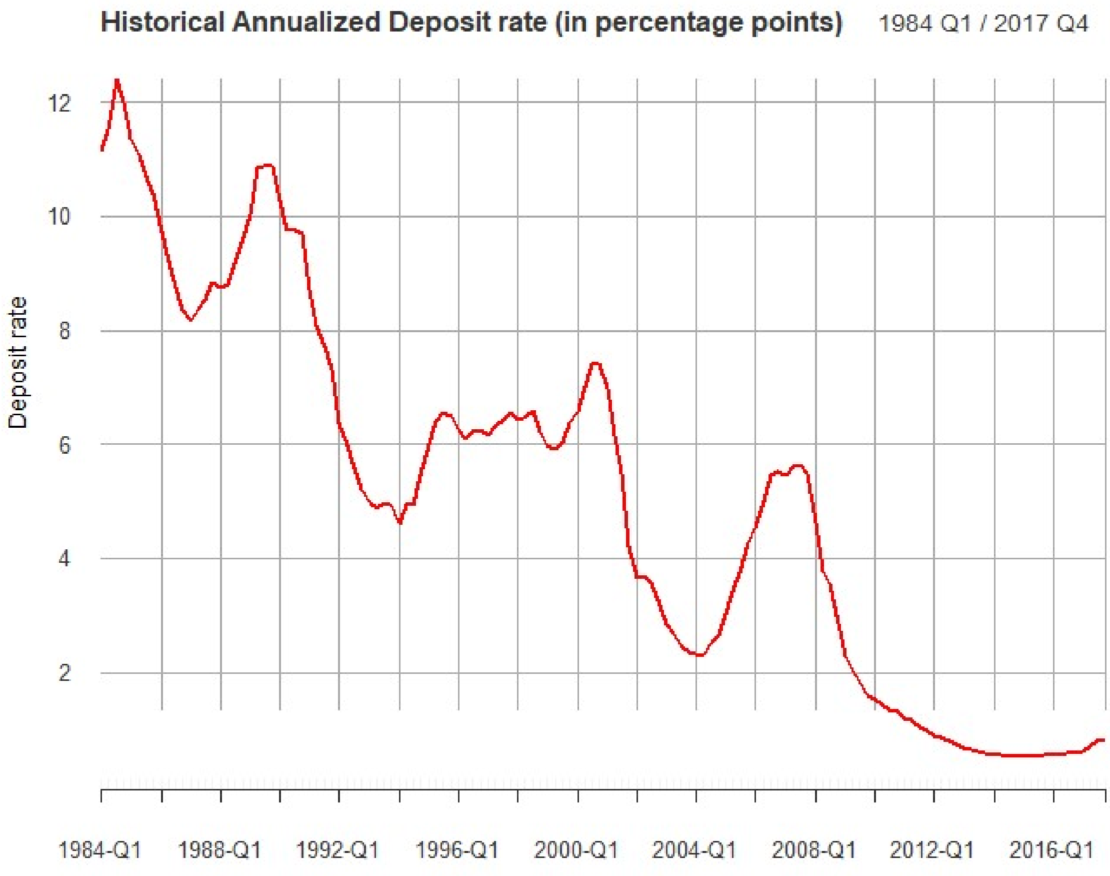
Domestic deposits non interest-bearing are not accounted for since these deposits do not generate expense for the banks, hence should not impact average deposit rate.

It is also assumed that all foreign deposits generate pay interest to the clients.

In the later stage, the quarterly historical deposite rate is annualized in order to be consistent with historical US Treasury Yields:

*Annual Deposit Rate(t) = 100 * ((1 + Deposit Rate(t)) ^ (4) – 1)*

Figure 2 Historical annualized quarterly deposite rate paid



Alternative approaches (given more data and time would be available):

- Split Interest paid into US and foreign accounts, calculate two separate deposit rates – it would probably better reflect differences between US domestic and foreign accounts (the model of later rate could probably include FX rates)

- Use only Foreign accounts paying interest (assuming that there are positive foreign accounts non interest bearing)

c): (Bonus) *Please write a gradient descent algorithm in python or R to arrive at the coefficients for below linear regression:*
<Deposit Rate Paid ~ TSY 1 Month>

The gradient descent algorithm has been implemented in *solution.R* script. The following values have been set while initializing the algorithm:

Table 1 Initial parameters in gradient descent

Parameter	Value
Theta0 (intercept)	0
Theta1 (coefficient TSY 1 Month)	1
Learning rate (alpha)	0.1

In order to verify the results of the gradient descent algorithm, the final value of coefficients has been compared to output of OLS fit with lm() function:

Table 2 Comparison of linear regression and PCA results

Parameter	Gradient descent	Lm() function	Percentage difference
Theta0 (intercept)	1.03514269	1.035159	-0.00157560%
Theta1 (coefficient TSY 1 Month)	0.9994190	0.999413	0.0009005313%

Q2 Modeling*: by using the treasury rate, build a model to describe the deposit rate.*

In this task, I investigate two modelling approaches:

1. **OLS (Ordinary Least Squares)**
   The first method relies on searching for linear relationships between dependent (deposit rate) and independent (US Treasury Yields) variables. Although it can be argued that macroeconomic relationships are in fact more complex than linear, this approach helps finding, in a relatively simple manner, the main links between macro factors and deposit rate.

   One of the main assumptions behind OLS states that both dependent variable (Y) and independent variables (X) are stationary processes. In case this assumption is violated, the regression can often lead to spurious regression, unless the two processes are not cointegrated. Figure 1 and Figure 2 indicate that both US Treasury Yields and Deposit Rate have unit root (visible time trends, covariance not constant in time). In order to verify this statement formally, ADF tests have been run on the historical time series:

Table 3 Augmented Dickey-Fuller test results for dependent and independent variables

Variable	ADF Test p-value	KPSS Test p-value
Deposit Rate	<0.01	<0.01
US TY 1M	0.04	0.01562
US TY 3M	<0.01	<0.01
US TY 6M	0.0186	<0.01
US TY 1Y	0.2	<0.01
US TY 2Y	<0.01	<0.01
US TY 5Y	0.44	<0.01
US TY 10Y	0.52	<0.01
US TY 20Y	0.0128	<0.01
US TY 30Y	0.015	<0.01

Although ADF test results are inconclusive (for some variables p-value below 5% indicates lack of unit root, for those with high p-value, there is a high probability of unit-root presence), KPSS test results point strongly towards lack of time series stationarity. Given those results and visual presence of time trends in Figures 1 and 2 presented in first sections of the document, I decided to differentiate Y (dependent) and X (independent) variables, i.e. to regress Quarter-on-Quarter differences in Y on Quarter-on-Quarter differences of X.

The aim of this exercise is to determine which combination of explanatory variables best reflects the variation in Y. The approach taken begins with first regressing Y on all explanatory variables and iteratively removing from the equation the drivers with highest p-value (the least significant ones). The results of the analysis are presented below.

Table 4 Iterative process of macroeconomic factors exclusion from linear model

Idx	Formula	Adjusted R2	Variable with highest p-value
1	Y.QoQ ~ US.Y.1M0.QoQ + US.Y.1Y.QoQ + US.Y.2Y.QoQ + US.Y.3M0.QoQ + US.Y.5Y.QoQ + US.Y.6M0.QoQ + US.Y.7Y.QoQ + US.Y.10Y.QoQ + US.Y.20Y.QoQ + US.Y.30Y.QoQ	76.07%	US.Y.20Y.QoQ
2	Y.QoQ ~ US.Y.1M0.QoQ + US.Y.1Y.QoQ + US.Y.2Y.QoQ + US.Y.3M0.QoQ + US.Y.5Y.QoQ + US.Y.6M0.QoQ + US.Y.7Y.QoQ + US.Y.10Y.QoQ + US.Y.30Y.QoQ	76.67%	US.Y.10Y.QoQ
3	Y.QoQ ~ US.Y.1M0.QoQ + US.Y.1Y.QoQ + US.Y.2Y.QoQ + US.Y.3M0.QoQ + US.Y.5Y.QoQ + US.Y.6M0.QoQ + US.Y.7Y.QoQ + US.Y.30Y.QoQ	77.15%	US.Y.1Y.QoQ
4	Y.QoQ ~ US.Y.1M0.QoQ + US.Y.2Y.QoQ + US.Y.3M0.QoQ + US.Y.5Y.QoQ + US.Y.6M0.QoQ + US.Y.7Y.QoQ + US.Y.30Y.QoQ	77.47%	US.Y.1M0.QoQ
5	Y.QoQ ~ US.Y.2Y.QoQ + US.Y.3M0.QoQ + US.Y.5Y.QoQ + US.Y.6M0.QoQ + US.Y.7Y.QoQ + US.Y.30Y.QoQ	60.73%	US.Y.5Y.QoQ

6	Y.QoQ ~ US.Y.2Y.QoQ + US.Y.3M0.QoQ + US.Y.6M0.QoQ + US.Y.7Y.QoQ + US.Y.30Y.QoQ	60.35%	US.Y.30Y.QoQ
7	Y.QoQ ~ US.Y.2Y.QoQ + US.Y.3M0.QoQ + US.Y.6M0.QoQ + US.Y.7Y.QoQ	60.68%	US.Y.3M0.QoQ
8	Y.QoQ ~ US.Y.2Y.QoQ + US.Y.6M0.QoQ + US.Y.7Y.QoQ	60.42%	US.Y.7Y.QoQ
9	**Y.QoQ ~ US.Y.2Y.QoQ + US.Y.6M0.QoQ**	**59.84%**	

The iterative exclusion of insignificant (or marginally significant) variables led to a model driven by QoQ difference in 6M US TY and 2Y US TY. Although adjusted R2 for the final regression model is lower than in case of the model 1, it should be noted that the first model includes 10 highly correlated variables, also many of them having no significant explanatory power. Moreover, R2 measure, even with penalty for high number of variables included (adjustment), generally increases with higher number of explanatory variables, hence should be applied with additional metrics / criteria.

Below are presented summary of results for model 9:

```
Call:
lm(formula = "Y.QoQ ~ US.Y.2Y.QoQ + US.Y.6M0.QoQ", data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.71463 -0.12609  0.01185  0.09413  0.92846 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.04840    0.01937  -2.499   0.0137 *  
US.Y.2Y.QoQ  -0.56169    0.08497  -6.610 8.66e-10 ***
US.Y.6M0.QoQ  1.05424    0.09077  11.614  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2228 on 132 degrees of freedom
  (91 observations deleted due to missingness)
Multiple R-squared:  0.6043,    Adjusted R-squared:  0.5984 
F-statistic: 100.8 on 2 and 132 DF,  p-value: < 2.2e-16
```

The coefficient of QoQ difference in 6M US Treasury yield is significant and positive. This result is rather intuitive given that, with rising interest rates, banks offer higher rates on deposits for the clients, whereas in zero interest rate environment clients are paid zero or little interest on the deposits. It could be also interpreted as a level factor of interest rate term structure. The second coefficient of QoQ difference in 2Y US Treasury yield is significant and negative. This outcome could be viewed as slope factor of interest rate curve.

(In case of more time would be available), the following steps should be taken in the analysis:

- Run exhaustive search algorithm (e.g. implemented in R leaps package) to get more alternative model specifications along with other dependent variable transformations (logarithm, QoQ logarithm), other explanatory macroeconomic variables (e.g. US inflation, FX rates for foreign deposits, market volatility, US economic growth etc), add lag dependent variable(s) to search

- Run tests to verify if OLS assumptions have not been violated (stationarity of residuals, lack of autocorrelation, homoscedasticity, normality, a lack of multicollinearity, linearity in parameters)
- Check if any outlier could significantly affect linear relationship in parameters
- Check stability of relationship, e.g. by rerunning the model on different subsamples (iteratively backward and forward) to determine how much coefficients vary depending on the estimation subsample
- Run in-sample and out-of-sample backtesting to assess model performance
- Check how much the models are affected after exclusion from estimation sample stress periods (e.g. 1999/2000, 2008/2009) to confirm whether the relationship relies only on few stress periods or is stable across the sample (perhaps also cross validate the regression)
- Deeper research (literature on economic intuition of interest rate models)

2. PCA (Principal Component Analysis)

As briefly pointed out in the previous point, modelling deposit rate with multiple tenors of US Treasury Yields poses two challenges:

- There is a strong autocorrelation between the different tenors of the yield curve;
- Time-series of yields tend to have a unit-root, which creates a high risk of spurious regressions.

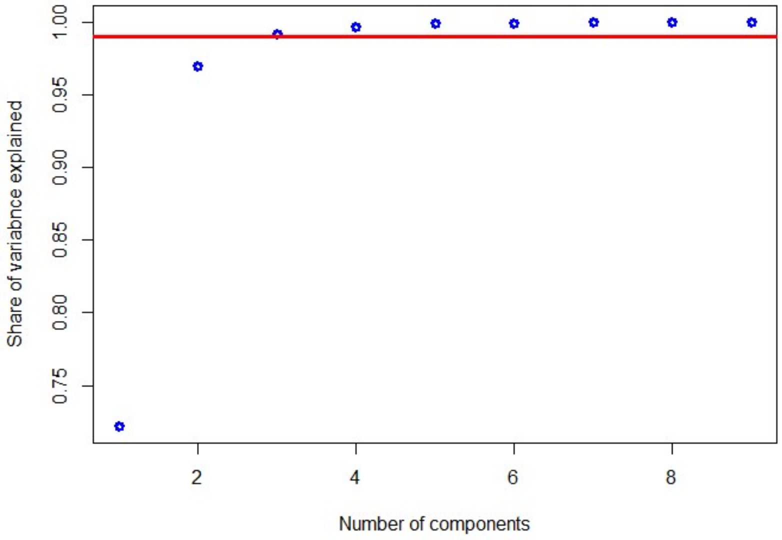
Although the second problem can be to some extent mitigated by differencing time series, the correlation of treasury yields is more difficult to handle via OLS. One solution is an application of principal component analysis (PCA). This dimensionality reduction method transforms correlated historical values into the factors independent of each other. Later on, based on the amount of variance explained, modeller can determine how many of the factors are to be applied in the further analysis.

Thanks to applying this method, the lower number of uncorrelated drivers can be applied in OLS. Also, in terms of interest rate modelling, the other advantage is that principal components can be interpreted as level, slope or curvature of interest rate curve.

In this exercise, similarly to OLS, both dependent and independent variables are transformed to QoQ differences. Due to a lack of data between 2002 and 2005, 30Y US Treasury Yield has been discarded from the sample. The PCA is run on the first 90% of data available for each variable (training set from 2001 Q4 to 2014 Q2) and further tested out of sample (testing set from 2014 Q3 to 2017 Q4).

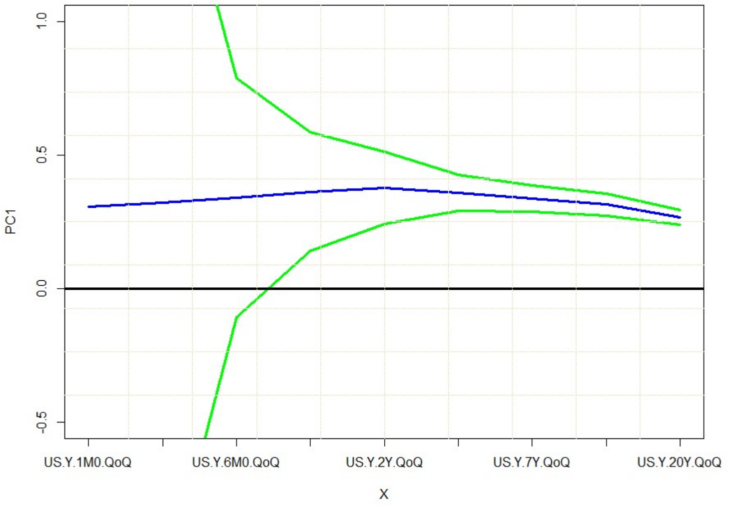
As presented in the figure below, first three principal components explain over 99% of variance. Since additional factors do not contribute significantly to improvement of model performance, the three components are selected.

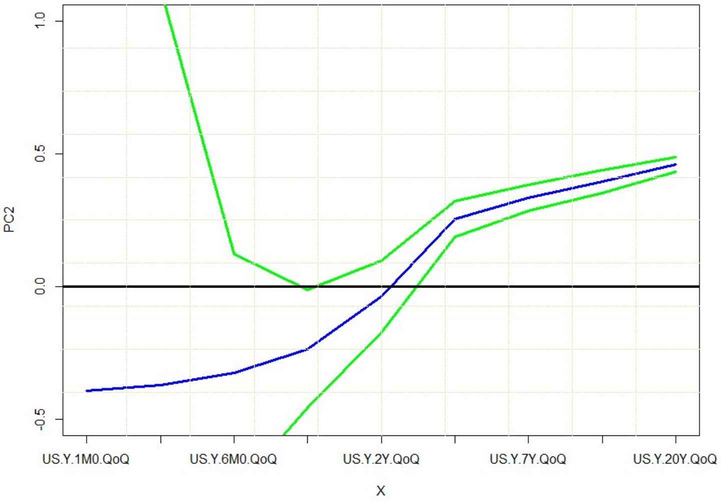
Figure 3 Share of variance explained dependent on number of principal components selected

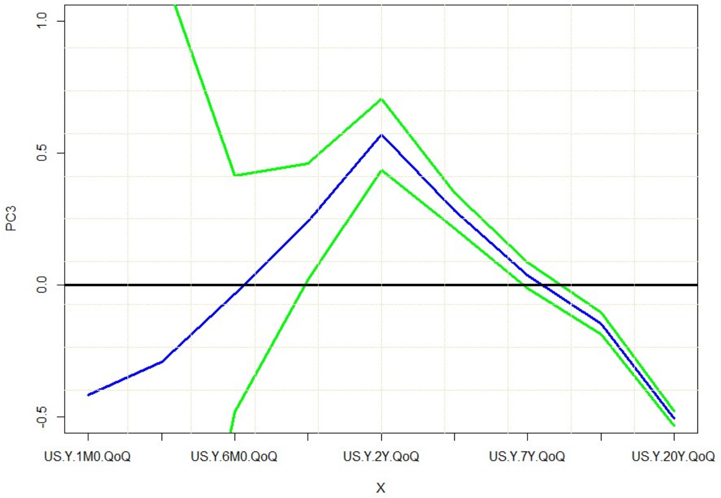


The following graphs present principal components with 1 standard deviations shocks dependent on US Treasury tenors. The first component could be interpreted as a change in level of interest rate curve, the second one as a change in slope and the third as a change in curvature (a more in depth economic explanation could be a part of further analysis).

Figure 4: Effects of 1 s.d. shocks on each principal component.







For comparison purposes between linear regression selected and PCA, I verified two statistics:

- Adjusted R2 in training sample (2001 Q4 – 2014 Q2)
- Out of sample RMSE in testing sample (2014 Q3 – 2017 Q4)

The actuals vs fitted plots in testing sets and a table with summary of results are depicted below.

Figure 5 Observed and predicted deposit rates (black) vs fitted values by OLS (blue) and PCA (red) in testing sample (2014 Q3 – 2017 Q4)

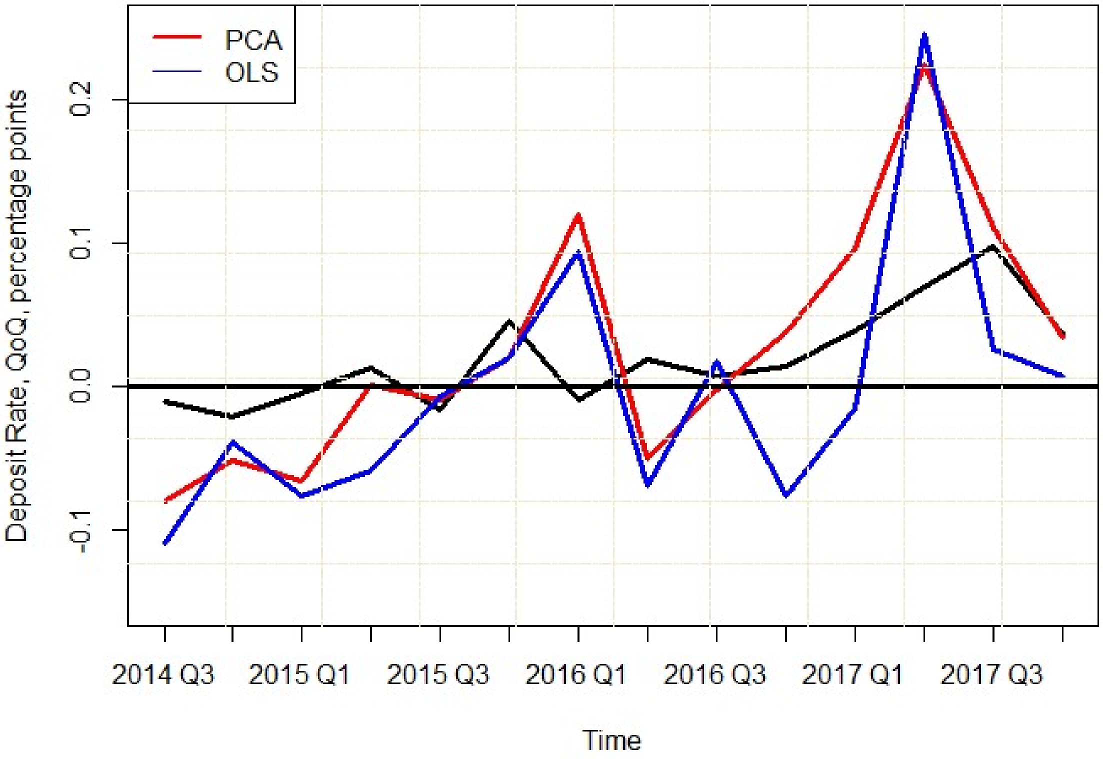


Table 5 Comparison of adjusted R2 and out-of-sample RMSE between linear regression and PCA

Model	Adjusted R2 (2001 Q2 – 2014 Q2)	Out of sample RMSE (2014 Q3 – 2017 Q4)
OLS (US TY 6M, 2Y)	73.17%	0.07794084
PCA (3 components)	69.08%	0.06572142

Adjusted R2 in training sample is higher in case of linear regression. However, when the model is used to predict out of sample deposit rates in the last 14 quarters, PCA yields closer fit to actuals. This finding could be interpreted as bias-variance trade off, meaning that model fitting closer to training sample (lower bias), is characterized by higher error out of sample (higher variance).

Given more intuitive interpretation of results and better results out of sample, I would rather select PCA for further analysis to improve its performance.

It should be noted, however, that there is still room for improvement. Both models produced rather volatile output out of sample compared to actuals. Also, the prediction in the last 6 quarters seems to be rather aggressive, whereas in stress testing framework conservativeness of results is a sought-after model property. Below are enlisted points that should be considered in further analysis.

Areas to be investigated:

- Application of other ML algorithms (Lasso regression, Decision Trees, Random Forrest, Support Vector Regression, Neural Network Regression, KNN)
- Application of different econometric models (e.g ARIMA, Error Correction Model)
- Application of higher number of evaluation measures, e.g. Mean Square Error, Mean Absolute Error, k-fold cross validation, Bayesian Information Criterion,
- Measuring performance on detransformed variable (Deposit Rate paid in levels instead of QoQ changes)
- Inclusion of other macroeconomic variables in the exercise, as provided in section on OLS