EXAMINATION QUESTION PAPER - Course paper

EBA 35301
Causality, Machine learning and Forecasting

Department of Economics

Start date:	08.03.2021	Time 09.00
Finish date:	15.03.2021	Time 12.00

Weight:	25% of EBA 3530
Total no. of pages:	4 incl. front page
No. of attachments files to question paper:	0
To be answered:	In groups of 1 - 3 students.

Answer paper size:	No restrictions excl. attachments
Max no. of answer paper attachment files:	1
Allowed answer paper file types:	pdf
Allowed answer paper attachment file types:	docx, ipynb, jpg, pdf, txt

EBA3530 – Group assignment 2021

Instructions

Prepare your answers as one group. The answers should be short and written together with the codes into a Jupyter notebook. Declare which version of Python you have used when preparing the answers.

The score for this assignment counts for 25% of the final grade. We use the scale 0-100 for grading.

Good luck!

Assignment

A) Regression analysis and Monte Carlo

Use the data labelled "randomX.csv" in the "Home Assignment" folder on Itslearning. Assume a true DGP like:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \varepsilon_i$$

where the errors are drawn from the normal distribution with a variance of 1. In this exercise you will conduct a Monte Carlo (MC) experiment. First, simulate 2,000 MC samples of Y_i using $\beta_0 = 0.5$ and $\beta_1 = 1.5$. On each of the MC samples, estimate the following model with OLS:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{3,i}^2 + \varepsilon_i$$

i) Plot a histogram for each of the estimated parameters as well as the R-squared. Into your histograms, also plot the corresponding median (of the respective parameter estimate/R-squared).

ii) Construct 95% confidence intervals for each parameter based on the simulations. Which of the coefficients are significant?

iii) Redo the Monte Carlo experiment, but now estimate the true model structure for each simulation. Make a histogram of the β_1 estimate. How does it compare to the one you produced in A.i? Do you find a difference? Provide an explanation why/why not.

iv) Explain in which context it would be useful to replace the OLS estimator used above with a logistic regression.

v) What is the interpretation of the β_2 coefficient above? What is the interpretation of β_3 and β_4?
Hint: β_3 and β_4 should only be interpreted together. It might help you to first derive an expression for the effect of a unit increase in X_3 on Y.

vi) Write shortly about how changing the number of Monte Carlo simulations and the number of observations (N) might affect your results in these experiments.

B) Time series

Use the data labelled "varekonsum.xlsx" in the "Home Assignment" folder on Itslearning. This data contains an index of household consumption of goods in Norway.

i. Load and plot the data. By just looking at the data, what time series properties would you say are prominent for this data?
Hint: To load the data use the below code.

```
# read in the data sheet
varekonsum = pd.read_excel('varekonsum.xlsx', usecols="C:IT", header=3).T
varekonsum = varekonsum.dropna(axis=1)

# adjust columnlabel
varekonsum.columns =['varekonsum']

# replace the index with a machine readable format
varekonsum.index = pd.date_range(start='31/01/2000', end='31/12/2020', freq='M')
```

ii. Construct a dummy variable which has zeros for each month in the sample, but a one for each December month. Call this variable "dum".

iii. Run the following OLS regression:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 dum + \varepsilon_t$$

and compute and plot $Y_t - \hat{Y}_t = \varepsilon_t$. What time series properties would you say are prominent for ε_t? Discuss your answer in relation to question 2.i. Why is there a difference?

iv. Compute and plot $log(Y_t) - log(Y_{t-12})$, i.e., the year-on-year growth in the data. Call this variable "d12y". Plot the series. What time series properties would you say are prominent for d12y? Does the data look stationary and does it have a seasonal component? Explain your answer.

v. Estimate an AR(1) model for d12y. Is the AR model stationary?
Use the BIC to find the optimal lag length for the AR(p) model for d12y. What do you get?
Hint: If you use the OLS class from the lectures/tutorial, compute the BIC using the formula $BIC = n \cdot ln(\widehat{\sigma_\varepsilon^2}) + k \cdot ln(n)$

where "k" is the number of variables + 2.

C) Regularization and dimension reduction

Use the data labelled "2018_Financial_Data.csv" In the "Home Assignment" folder on Itslearning. Let "Dividend payments" be the outcome variable and use the remaining variables as predictors.

Hint: Set the company names as index and convert "sector" to a categorical variable (see the code snippet below). Also, replace missing values with 0.

```
# set company name as index
financial_data = financial_data.set_index('Unnamed: 0')

# convert sector to categorical
financial_data["Sector"] = financial_data["Sector"].astype('category')
financial_data["Sector"] = financial_data["Sector"].cat.codes
```

i. Explain why estimating a OLS regression will fail in this case
ii. Run a LASSO regression where you choose the penalization parameter using 5-fold cross validation.
iii. Explain shortly why cross validation is used in C.ii. Explain shortly why cross validation, and out-of-sample testing in general, is used to avoid in-sample over-fitting.
iv. Make a table showing the regression output. Which predictors are chosen? What is the R-squared of this regression?
v. Use all predictors and compute 1 common component using PCA. Plot this common component.
vi. Regress the outcome variable on the common component. What is the R-squared of this regression? Compare your result with what you got in C.iv.
vii. In light of your answer in C.vi, discuss shortly in which cases using PCA versus LASSO might be beneficial.