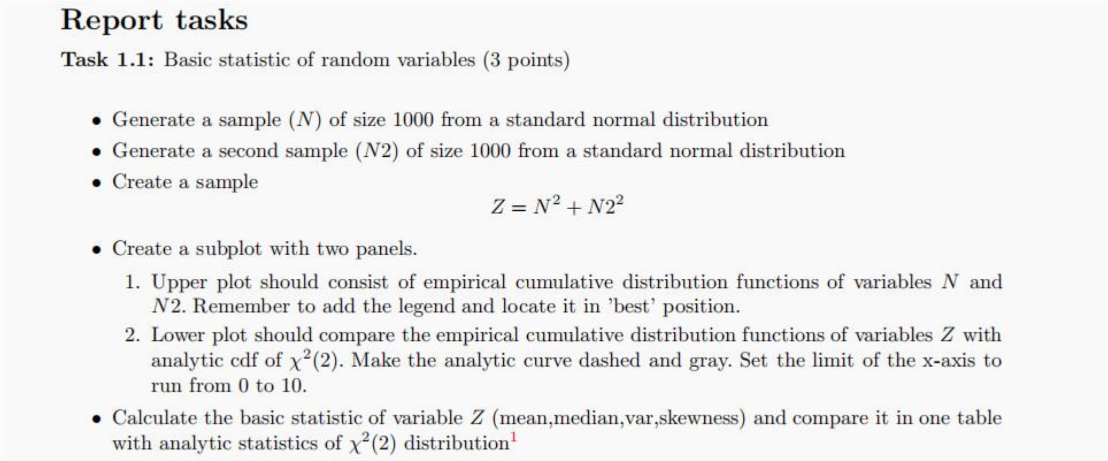
Report task list:

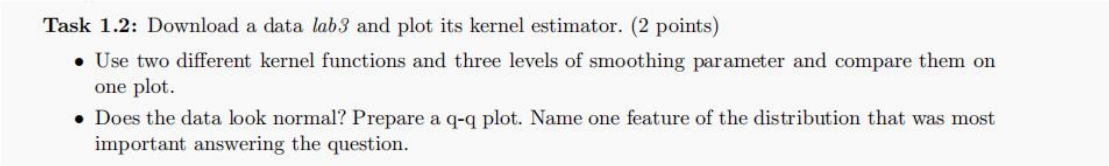
Report 1



Report tasks

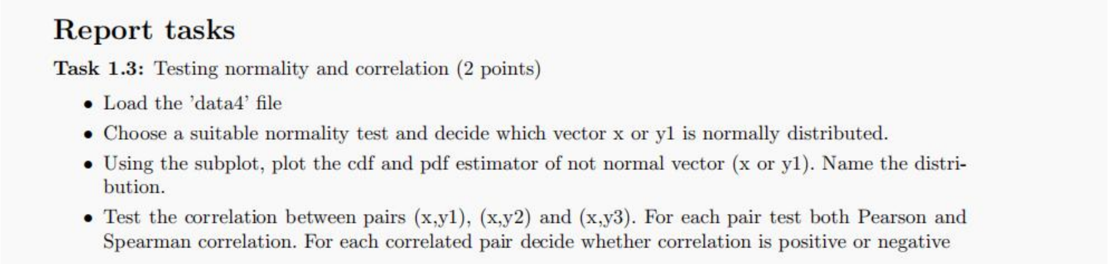
Task 1.1: Basic statistic of random variables (3 points)

- Generate a sample (N) of size 1000 from a standard normal distribution
- Generate a second sample ($N2$) of size 1000 from a standard normal distribution
- Create a sample
$$Z = N^2 + N2^2$$
- Create a subplot with two panels.
  1. Upper plot should consist of empirical cumulative distribution functions of variables N and $N2$. Remember to add the legend and locate it in 'best' position.
  2. Lower plot should compare the empirical cumulative distribution functions of variables Z with analytic cdf of $\chi^2(2)$. Make the analytic curve dashed and gray. Set the limit of the x-axis to run from 0 to 10.
- Calculate the basic statistic of variable Z (mean,median,var,skewness) and compare it in one table with analytic statistics of $\chi^2(2)$ distribution[1]



Task 1.2: Download a data *lab3* and plot its kernel estimator. (2 points)

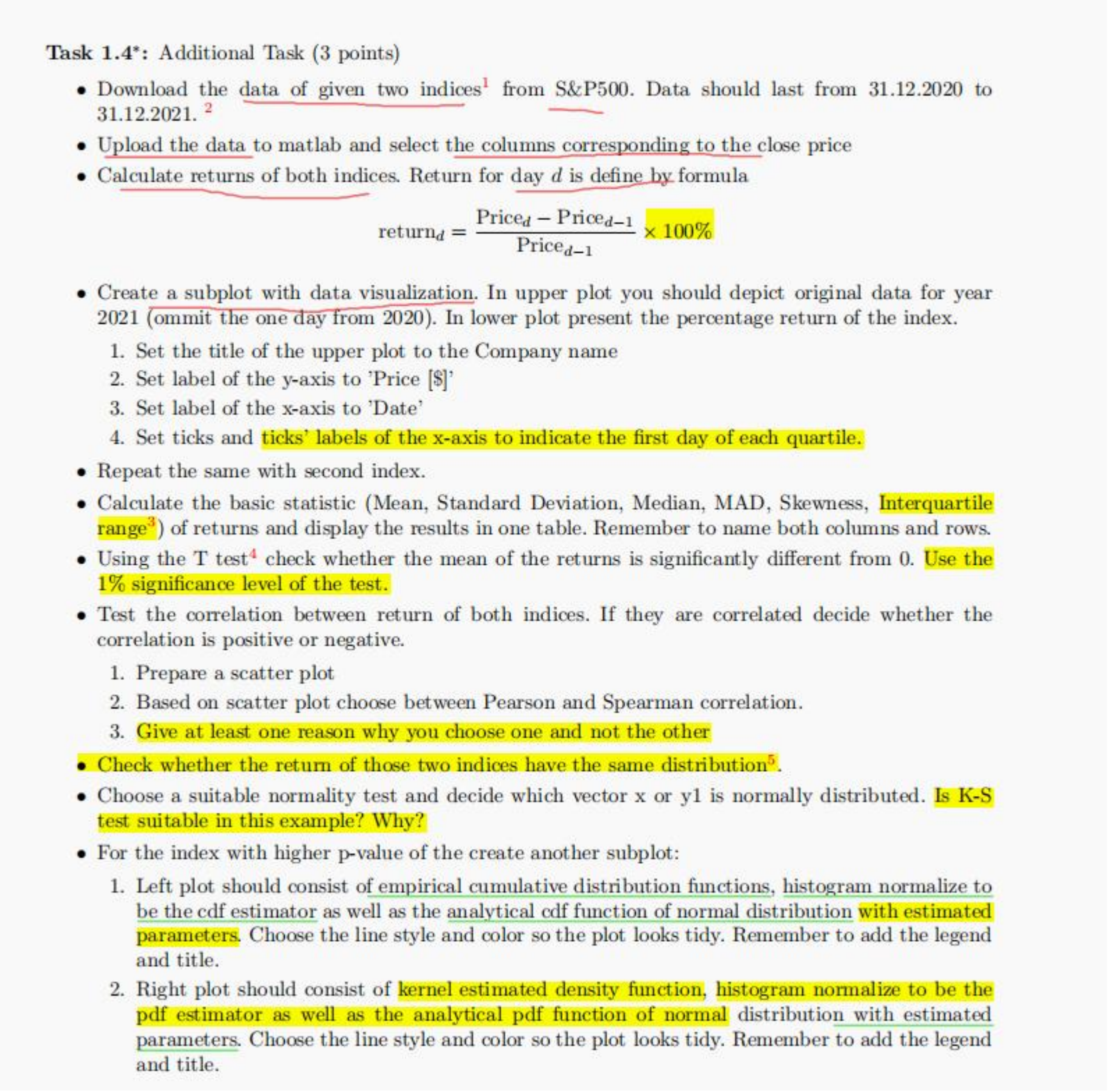
- Use two different kernel functions and three levels of smoothing parameter and compare them on one plot.
- Does the data look normal? Prepare a q-q plot. Name one feature of the distribution that was most important answering the question.



Report tasks

Task 1.3: Testing normality and correlation (2 points)

- Load the 'data4' file
- Choose a suitable normality test and decide which vector x or y1 is normally distributed.
- Using the subplot, plot the cdf and pdf estimator of not normal vector (x or y1). Name the distribution.
- Test the correlation between pairs (x,y1), (x,y2) and (x,y3). For each pair test both Pearson and Spearman correlation. For each correlated pair decide whether correlation is positive or negative



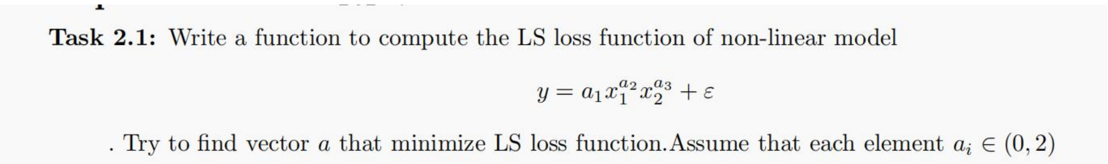
Task 1.4*: Additional Task (3 points)

- Download the data of given two indices[1] from S&P500. Data should last from 31.12.2020 to 31.12.2021. [2]
- Upload the data to matlab and select the columns corresponding to the close price
- Calculate returns of both indices. Return for day d is define by formula

$$\text{return}_d = \frac{\text{Price}_d - \text{Price}_{d-1}}{\text{Price}_{d-1}} \times 100\%$$

- Create a subplot with data visualization. In upper plot you should depict original data for year 2021 (ommit the one day from 2020). In lower plot present the percentage return of the index.
  1. Set the title of the upper plot to the Company name
  2. Set label of the y-axis to 'Price [$]'
  3. Set label of the x-axis to 'Date'
  4. Set ticks and ticks' labels of the x-axis to indicate the first day of each quartile.
- Repeat the same with second index.
- Calculate the basic statistic (Mean, Standard Deviation, Median, MAD, Skewness, Interquartile range[3]) of returns and display the results in one table. Remember to name both columns and rows.
- Using the T test[4] check whether the mean of the returns is significantly different from 0. Use the 1% significance level of the test.
- Test the correlation between return of both indices. If they are correlated decide whether the correlation is positive or negative.
  1. Prepare a scatter plot
  2. Based on scatter plot choose between Pearson and Spearman correlation.
  3. Give at least one reason why you choose one and not the other
- Check whether the return of those two indices have the same distribution[5].
- Choose a suitable normality test and decide which vector x or y1 is normally distributed. Is K-S test suitable in this example? Why?
- For the index with higher p-value of the create another subplot:
  1. Left plot should consist of empirical cumulative distribution functions, histogram normalize to be the cdf estimator as well as the analytical cdf function of normal distribution with estimated parameters. Choose the line style and color so the plot looks tidy. Remember to add the legend and title.
  2. Right plot should consist of kernel estimated density function, histogram normalize to be the pdf estimator as well as the analytical pdf function of normal distribution with estimated parameters. Choose the line style and color so the plot looks tidy. Remember to add the legend and title.

Report 2



Task 2.1: Write a function to compute the LS loss function of non-linear model

$$y = a_1 x_1^{a_2} x_2^{a_3} + \varepsilon$$

. Try to find vector a that minimize LS loss function.Assume that each element $a_i \in (0,2)$

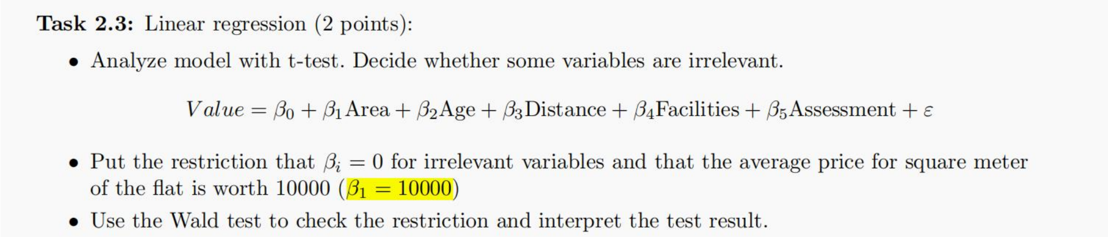
Task2.2

```
% Find and optimal model with 3 variables

% Use the t-test to check if in the optimal model consist of any irrelevant variable

%function about an autoregressive model.

%function about BIC
```

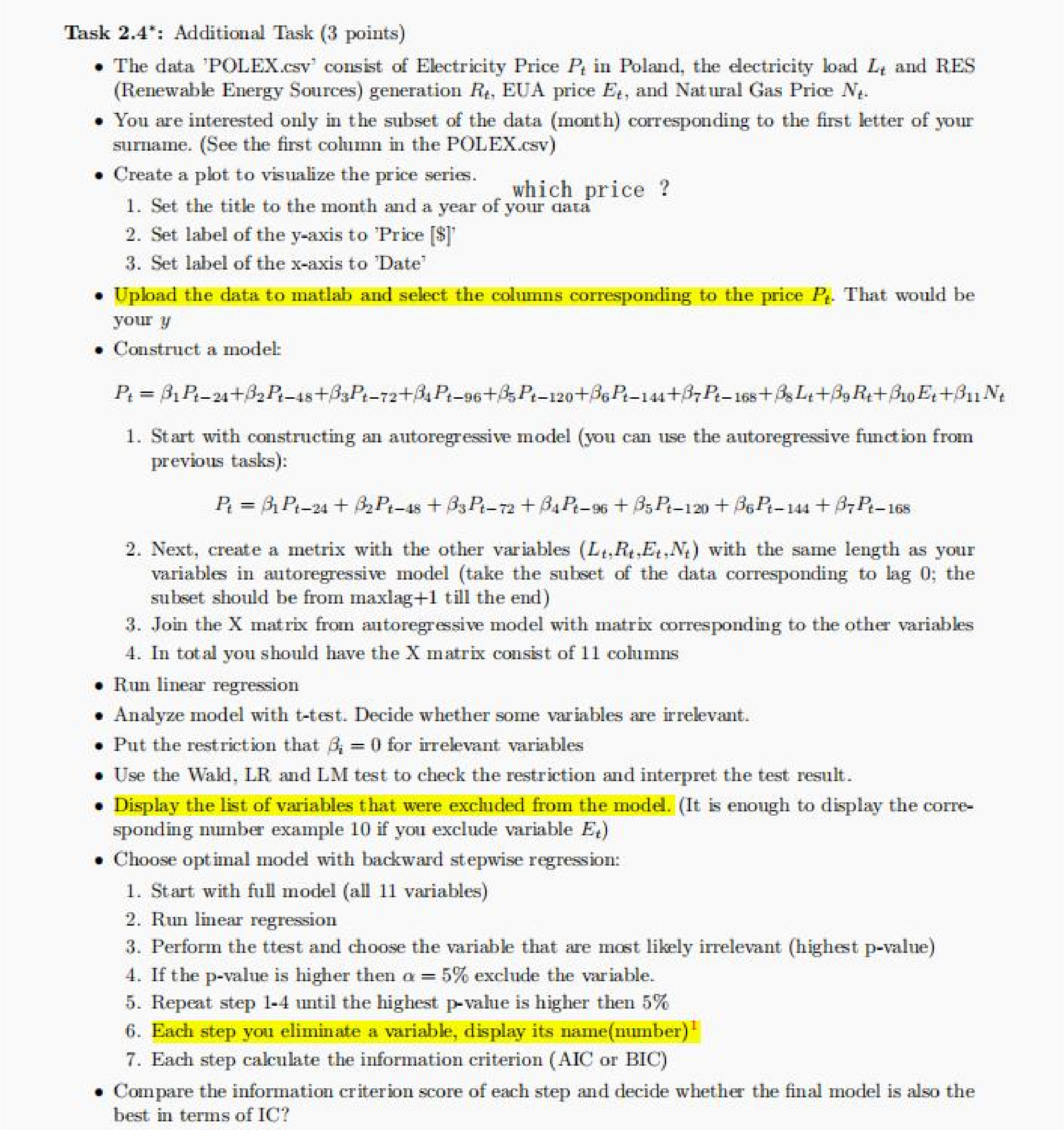


Task 2.3: Linear regression (2 points):

- Analyze model with t-test. Decide whether some variables are irrelevant.

$$Value = \beta_0 + \beta_1 \text{Area} + \beta_2 \text{Age} + \beta_3 \text{Distance} + \beta_4 \text{Facilities} + \beta_5 \text{Assessment} + \varepsilon$$

- Put the restriction that $\beta_i = 0$ for irrelevant variables and that the average price for square meter of the flat is worth 10000 ($\beta_1 = 10000$)
- Use the Wald test to check the restriction and interpret the test result.



Task 2.4*: Additional Task (3 points)

- The data 'POLEX.csv' consist of Electricity Price P_t in Poland, the electricity load L_t and RES (Renewable Energy Sources) generation R_t, EUA price E_t, and Natural Gas Price N_t.
- You are interested only in the subset of the data (month) corresponding to the first letter of your surname. (See the first column in the POLEX.csv)
- Create a plot to visualize the price series.

  which price ?

  1. Set the title to the month and a year of your data
  2. Set label of the y-axis to 'Price [$]'
  3. Set label of the x-axis to 'Date'
- Upload the data to matlab and select the columns corresponding to the price P_t. That would be your y
- Construct a model:

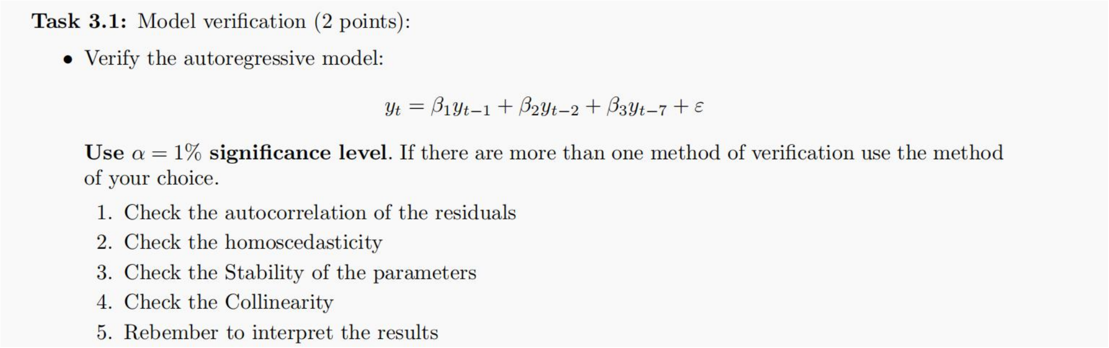
$$P_t = \beta_1 P_{t-24} + \beta_2 P_{t-48} + \beta_3 P_{t-72} + \beta_4 P_{t-96} + \beta_5 P_{t-120} + \beta_6 P_{t-144} + \beta_7 P_{t-168} + \beta_8 L_t + \beta_9 R_t + \beta_{10} E_t + \beta_{11} N_t$$

  1. Start with constructing an autoregressive model (you can use the autoregressive function from previous tasks):

$$P_t = \beta_1 P_{t-24} + \beta_2 P_{t-48} + \beta_3 P_{t-72} + \beta_4 P_{t-96} + \beta_5 P_{t-120} + \beta_6 P_{t-144} + \beta_7 P_{t-168}$$

  2. Next, create a metrix with the other variables (L_t,R_t,E_t,N_t) with the same length as your variables in autoregressive model (take the subset of the data corresponding to lag 0; the subset should be from maxlag+1 till the end)
  3. Join the X matrix from autoregressive model with matrix corresponding to the other variables
  4. In total you should have the X matrix consist of 11 columns
- Run linear regression
- Analyze model with t-test. Decide whether some variables are irrelevant.
- Put the restriction that $\beta_i = 0$ for irrelevant variables
- Use the Wald, LR and LM test to check the restriction and interpret the test result.
- Display the list of variables that were excluded from the model. (It is enough to display the corresponding number example 10 if you exclude variable E_t)
- Choose optimal model with backward stepwise regression:
  1. Start with full model (all 11 variables)
  2. Run linear regression
  3. Perform the ttest and choose the variable that are most likely irrelevant (highest p-value)
  4. If the p-value is higher then $\alpha = 5\%$ exclude the variable.
  5. Repeat step 1-4 until the highest p-value is higher then 5%
  6. Each step you eliminate a variable, display its name(number)[1]
  7. Each step calculate the information criterion (AIC or BIC)
- Compare the information criterion score of each step and decide whether the final model is also the best in terms of IC?

Report 3



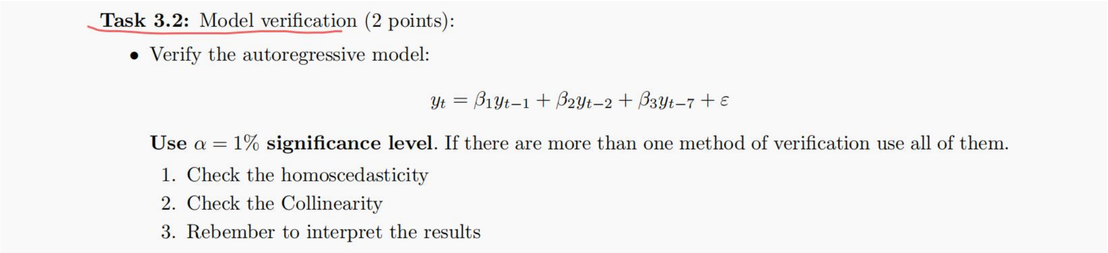
Task 3.1: Model verification (2 points):

- Verify the autoregressive model:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-7} + \varepsilon$$

Use $\alpha = 1\%$ **significance level**. If there are more than one method of verification use the method of your choice.

1. Check the autocorrelation of the residuals
2. Check the homoscedasticity
3. Check the Stability of the parameters
4. Check the Collinearity
5. Rebember to interpret the results



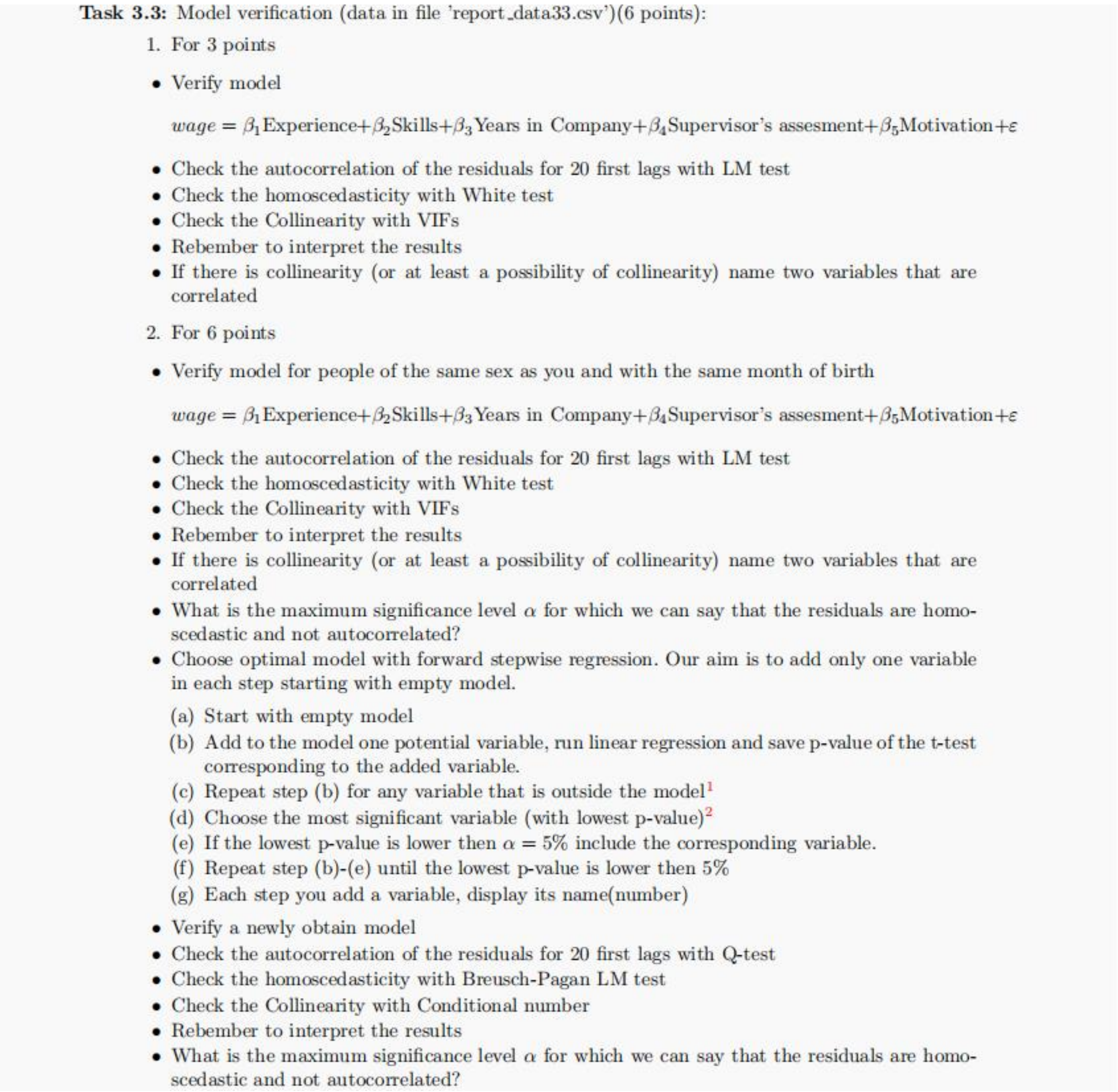
Task 3.2: Model verification (2 points):

- Verify the autoregressive model:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-7} + \varepsilon$$

Use $\alpha = 1\%$ **significance level**. If there are more than one method of verification use all of them.

1. Check the homoscedasticity
2. Check the Collinearity
3. Rebember to interpret the results



Task 3.3: Model verification (data in file 'report_data33.csv')(6 points):

1. For 3 points

- Verify model

  $wage = \beta_1$Experience$+\beta_2$Skills$+\beta_3$Years in Company$+\beta_4$Supervisor's assesment$+\beta_5$Motivation$+\varepsilon$

- Check the autocorrelation of the residuals for 20 first lags with LM test
- Check the homoscedasticity with White test
- Check the Collinearity with VIFs
- Rebember to interpret the results
- If there is collinearity (or at least a possibility of collinearity) name two variables that are correlated

2. For 6 points

- Verify model for people of the same sex as you and with the same month of birth

  $wage = \beta_1$Experience$+\beta_2$Skills$+\beta_3$Years in Company$+\beta_4$Supervisor's assesment$+\beta_5$Motivation$+\varepsilon$

- Check the autocorrelation of the residuals for 20 first lags with LM test
- Check the homoscedasticity with White test
- Check the Collinearity with VIFs
- Rebember to interpret the results
- If there is collinearity (or at least a possibility of collinearity) name two variables that are correlated
- What is the maximum significance level α for which we can say that the residuals are homoscedastic and not autocorrelated?
- Choose optimal model with forward stepwise regression. Our aim is to add only one variable in each step starting with empty model.

  (a) Start with empty model

  (b) Add to the model one potential variable, run linear regression and save p-value of the t-test corresponding to the added variable.

  (c) Repeat step (b) for any variable that is outside the model[1]

  (d) Choose the most significant variable (with lowest p-value)[2]

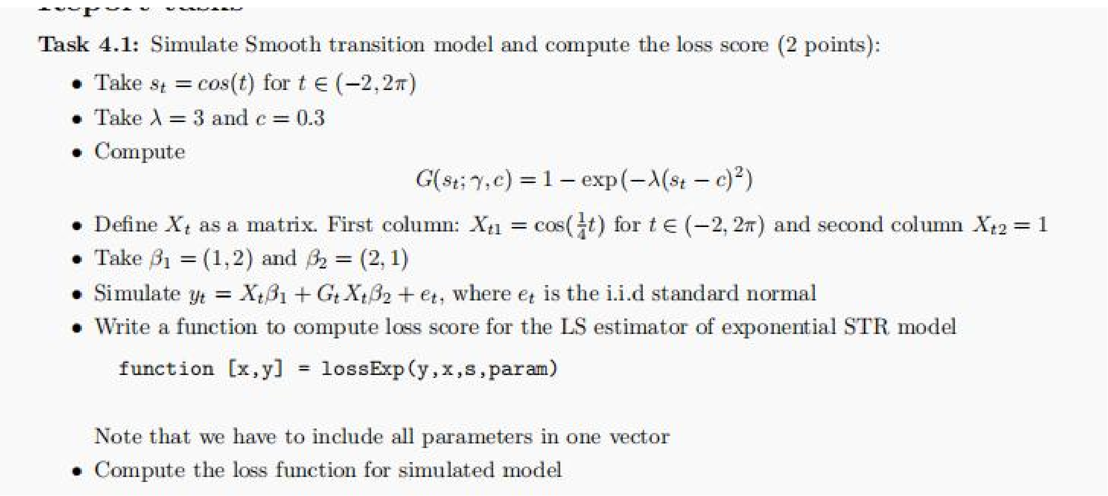
  (e) If the lowest p-value is lower then $\alpha = 5\%$ include the corresponding variable.

  (f) Repeat step (b)-(e) until the lowest p-value is lower then 5%

  (g) Each step you add a variable, display its name(number)

- Verify a newly obtain model
- Check the autocorrelation of the residuals for 20 first lags with Q-test
- Check the homoscedasticity with Breusch-Pagan LM test
- Check the Collinearity with Conditional number
- Rebember to interpret the results
- What is the maximum significance level α for which we can say that the residuals are homoscedastic and not autocorrelated?

Report 4



Task 4.1: Simulate Smooth transition model and compute the loss score (2 points):

- Take $s_t = cos(t)$ for $t \in (-2, 2\pi)$
- Take $\lambda = 3$ and $c = 0.3$
- Compute

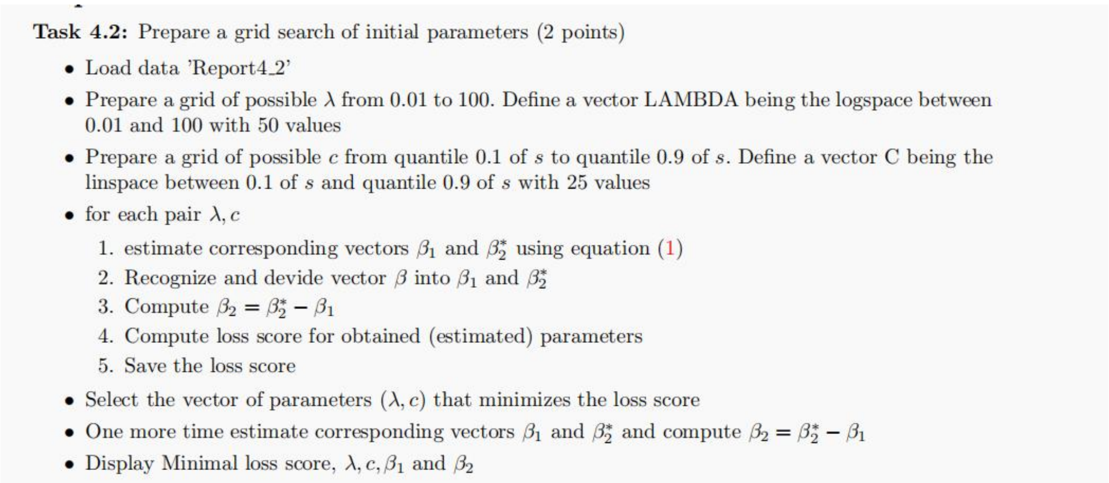
$$G(s_t; \gamma, c) = 1 - \exp(-\lambda(s_t - c)^2)$$

- Define X_t as a matrix. First column: $X_{t1} = \cos(\frac{1}{4}t)$ for $t \in (-2, 2\pi)$ and second column $X_{t2} = 1$
- Take $\beta_1 = (1, 2)$ and $\beta_2 = (2, 1)$
- Simulate $y_t = X_t\beta_1 + G_t X_t \beta_2 + e_t$, where e_t is the i.i.d standard normal
- Write a function to compute loss score for the LS estimator of exponential STR model

```
function [x,y] = lossExp(y,x,s,param)
```

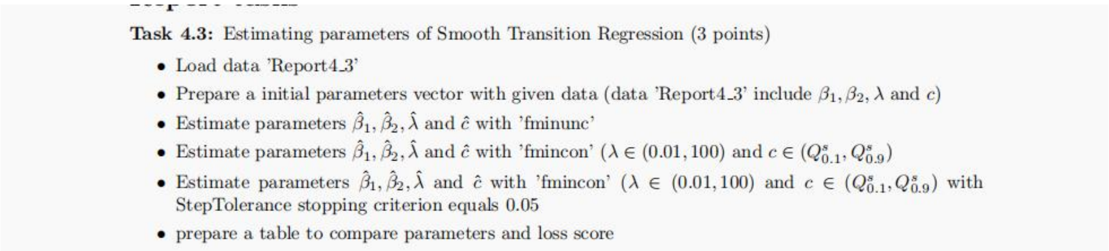
  Note that we have to include all parameters in one vector

- Compute the loss function for simulated model



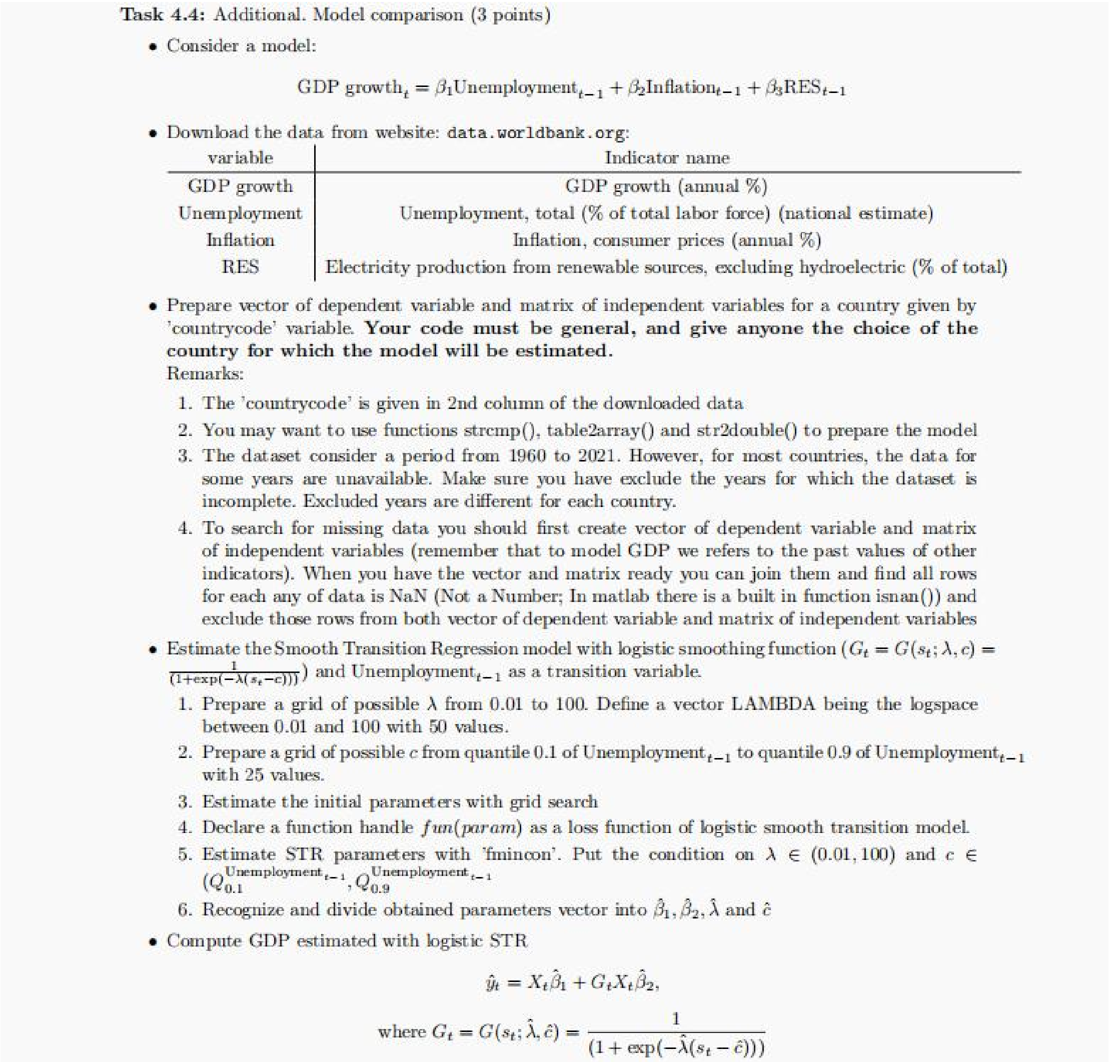
Task 4.2: Prepare a grid search of initial parameters (2 points)

- Load data 'Report4_2'
- Prepare a grid of possible λ from 0.01 to 100. Define a vector LAMBDA being the logspace between 0.01 and 100 with 50 values
- Prepare a grid of possible c from quantile 0.1 of s to quantile 0.9 of s. Define a vector C being the linspace between 0.1 of s and quantile 0.9 of s with 25 values
- for each pair λ, c
  1. estimate corresponding vectors β_1 and β_2^* using equation (1)
  2. Recognize and devide vector β into β_1 and β_2^*
  3. Compute $\beta_2 = \beta_2^* - \beta_1$
  4. Compute loss score for obtained (estimated) parameters
  5. Save the loss score
- Select the vector of parameters (λ, c) that minimizes the loss score
- One more time estimate corresponding vectors β_1 and β_2^* and compute $\beta_2 = \beta_2^* - \beta_1$
- Display Minimal loss score, λ, c, β_1 and β_2



Task 4.3: Estimating parameters of Smooth Transition Regression (3 points)

- Load data 'Report4_3'
- Prepare a initial parameters vector with given data (data 'Report4_3' include $\beta_1, \beta_2, \lambda$ and c)
- Estimate parameters $\hat{\beta}_1, \hat{\beta}_2, \hat{\lambda}$ and $\hat{c}$ with 'fminunc'
- Estimate parameters $\hat{\beta}_1, \hat{\beta}_2, \hat{\lambda}$ and $\hat{c}$ with 'fmincon' ($\lambda \in (0.01, 100)$ and $c \in (Q^s_{0.1}, Q^s_{0.9})$)
- Estimate parameters $\hat{\beta}_1, \hat{\beta}_2, \hat{\lambda}$ and $\hat{c}$ with 'fmincon' ($\lambda \in (0.01, 100)$ and $c \in (Q^s_{0.1}, Q^s_{0.9})$ with StepTolerance stopping criterion equals 0.05
- prepare a table to compare parameters and loss score



Task 4.4: Additional. Model comparison (3 points)

- Consider a model:

$$\text{GDP growth}_t = \beta_1 \text{Unemployment}_{t-1} + \beta_2 \text{Inflation}_{t-1} + \beta_3 \text{RES}_{t-1}$$

- Download the data from website: `data.worldbank.org`:

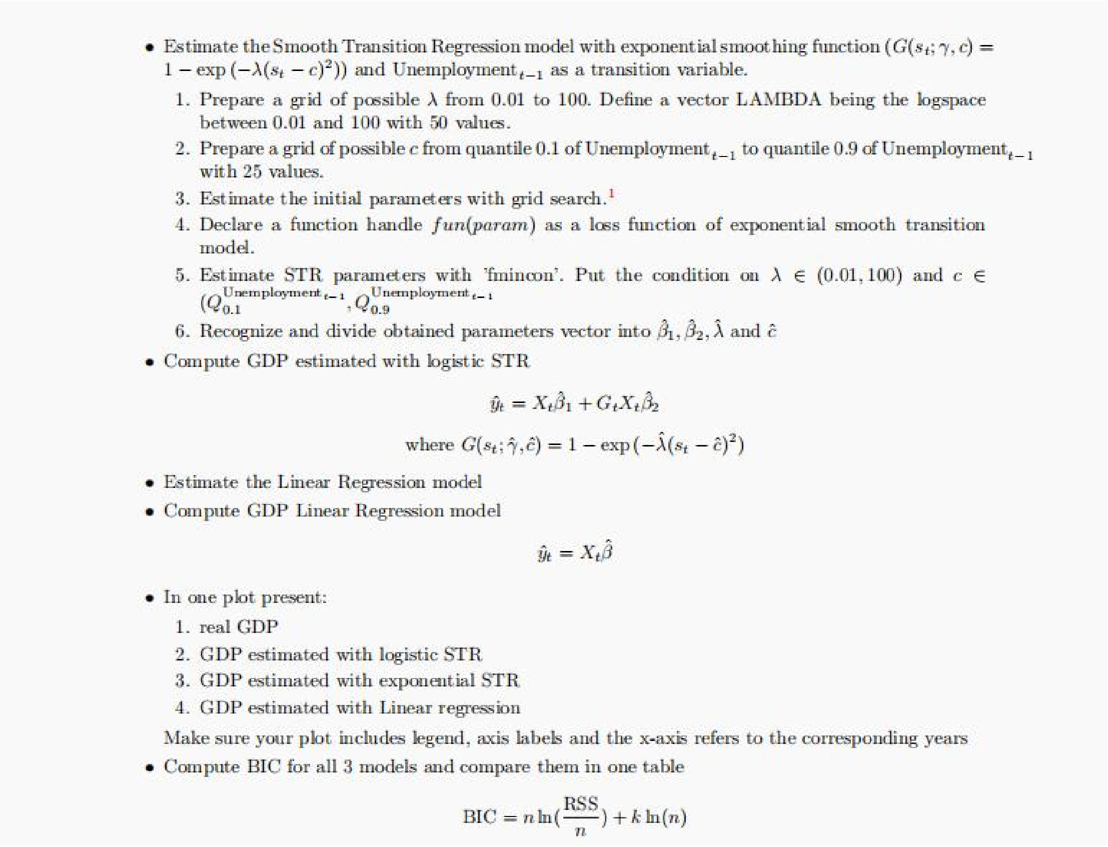
variable	Indicator name
GDP growth	GDP growth (annual %)
Unemployment	Unemployment, total (% of total labor force) (national estimate)
Inflation	Inflation, consumer prices (annual %)
RES	Electricity production from renewable sources, excluding hydroelectric (% of total)

- Prepare vector of dependent variable and matrix of independent variables for a country given by 'countrycode' variable. **Your code must be general, and give anyone the choice of the country for which the model will be estimated.**

  Remarks:
  1. The 'countrycode' is given in 2nd column of the downloaded data
  2. You may want to use functions strcmp(), table2array() and str2double() to prepare the model
  3. The dataset consider a period from 1960 to 2021. However, for most countries, the data for some years are unavailable. Make sure you have exclude the years for which the dataset is incomplete. Excluded years are different for each country.
  4. To search for missing data you should first create vector of dependent variable and matrix of independent variables (remember that to model GDP we refers to the past values of other indicators). When you have the vector and matrix ready you can join them and find all rows for each any of data is NaN (Not a Number; In matlab there is a built in function isnan()) and exclude those rows from both vector of dependent variable and matrix of independent variables
- Estimate the Smooth Transition Regression model with logistic smoothing function ($G_t = G(s_t; \lambda, c) = \frac{1}{(1+\exp(-\lambda(s_t - c)))}$) and $\text{Unemployment}_{t-1}$ as a transition variable.
  1. Prepare a grid of possible λ from 0.01 to 100. Define a vector LAMBDA being the logspace between 0.01 and 100 with 50 values.
  2. Prepare a grid of possible c from quantile 0.1 of $\text{Unemployment}_{t-1}$ to quantile 0.9 of $\text{Unemployment}_{t-1}$ with 25 values.
  3. Estimate the initial parameters with grid search
  4. Declare a function handle $fun(param)$ as a loss function of logistic smooth transition model.
  5. Estimate STR parameters with 'fmincon'. Put the condition on $\lambda \in (0.01, 100)$ and $c \in (Q_{0.1}^{\text{Unemployment}_{t-1}}, Q_{0.9}^{\text{Unemployment}_{t-1}}$
  6. Recognize and divide obtained parameters vector into $\hat{\beta}_1, \hat{\beta}_2, \hat{\lambda}$ and $\hat{c}$
- Compute GDP estimated with logistic STR

$$\hat{y}_t = X_t \hat{\beta}_1 + G_t X_t \hat{\beta}_2,$$

$$\text{where } G_t = G(s_t; \hat{\lambda}, \hat{c}) = \frac{1}{(1 + \exp(-\hat{\lambda}(s_t - \hat{c})))}$$



- Estimate the Smooth Transition Regression model with exponential smoothing function ($G(s_t;\gamma,c) = 1-\exp\left(-\lambda(s_t-c)^2\right)$) and Unemployment$_{t-1}$ as a transition variable.
  1. Prepare a grid of possible λ from 0.01 to 100. Define a vector LAMBDA being the logspace between 0.01 and 100 with 50 values.
  2. Prepare a grid of possible c from quantile 0.1 of Unemployment$_{t-1}$ to quantile 0.9 of Unemployment$_{t-1}$ with 25 values.
  3. Estimate the initial parameters with grid search.[1]
  4. Declare a function handle $fun(param)$ as a loss function of exponential smooth transition model.
  5. Estimate STR parameters with 'fmincon'. Put the condition on $\lambda \in (0.01, 100)$ and $c \in (Q_{0.1}^{\text{Unemployment}_{t-1}}, Q_{0.9}^{\text{Unemployment}_{t-1}}$
  6. Recognize and divide obtained parameters vector into $\hat{\beta}_1, \hat{\beta}_2, \hat{\lambda}$ and $\hat{c}$
- Compute GDP estimated with logistic STR

$$\hat{y}_t = X_t\hat{\beta}_1 + G_tX_t\hat{\beta}_2$$

$$\text{where } G(s_t;\hat{\gamma},\hat{c}) = 1-\exp\left(-\hat{\lambda}(s_t-\hat{c})^2\right)$$

- Estimate the Linear Regression model
- Compute GDP Linear Regression model

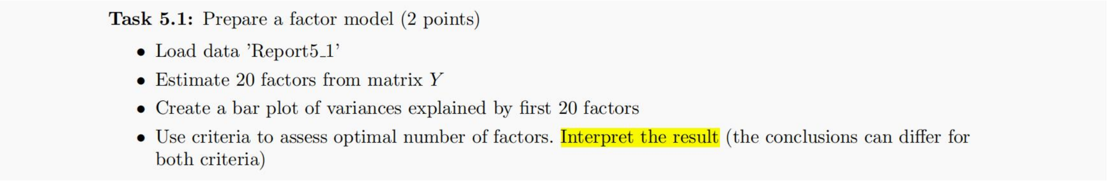
$$\hat{y}_t = X_t\hat{\beta}$$

- In one plot present:
  1. real GDP
  2. GDP estimated with logistic STR
  3. GDP estimated with exponential STR
  4. GDP estimated with Linear regression

  Make sure your plot includes legend, axis labels and the x-axis refers to the corresponding years
- Compute BIC for all 3 models and compare them in one table

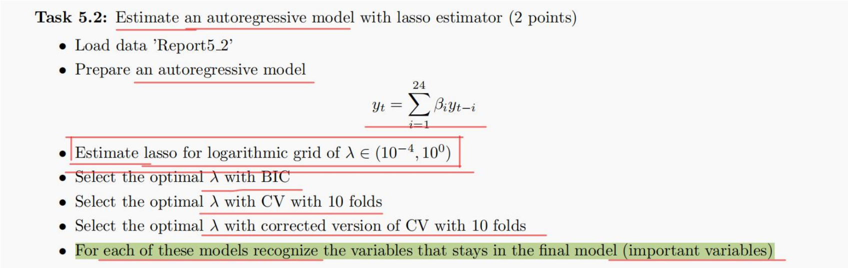
$$\text{BIC} = n\ln\left(\frac{\text{RSS}}{n}\right) + k\ln(n)$$

Report 5



Task 5.1: Prepare a factor model (2 points)

- Load data 'Report5_1'
- Estimate 20 factors from matrix Y
- Create a bar plot of variances explained by first 20 factors
- Use criteria to assess optimal number of factors. Interpret the result (the conclusions can differ for both criteria)

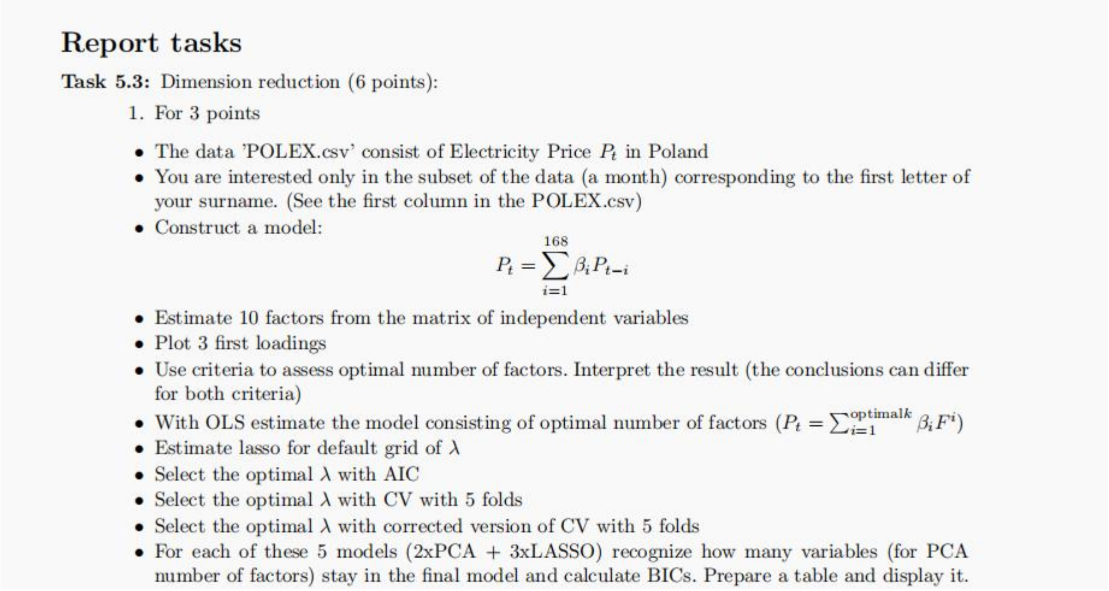


Task 5.2: Estimate an autoregressive model with lasso estimator (2 points)

- Load data 'Report5_2'
- Prepare an autoregressive model

$$y_t = \sum_{i=1}^{24} \beta_i y_{t-i}$$

- Estimate lasso for logarithmic grid of $\lambda \in (10^{-4}, 10^0)$
- Select the optimal λ with BIC
- Select the optimal λ with CV with 10 folds
- Select the optimal λ with corrected version of CV with 10 folds
- For each of these models recognize the variables that stays in the final model (important variables)



Report tasks

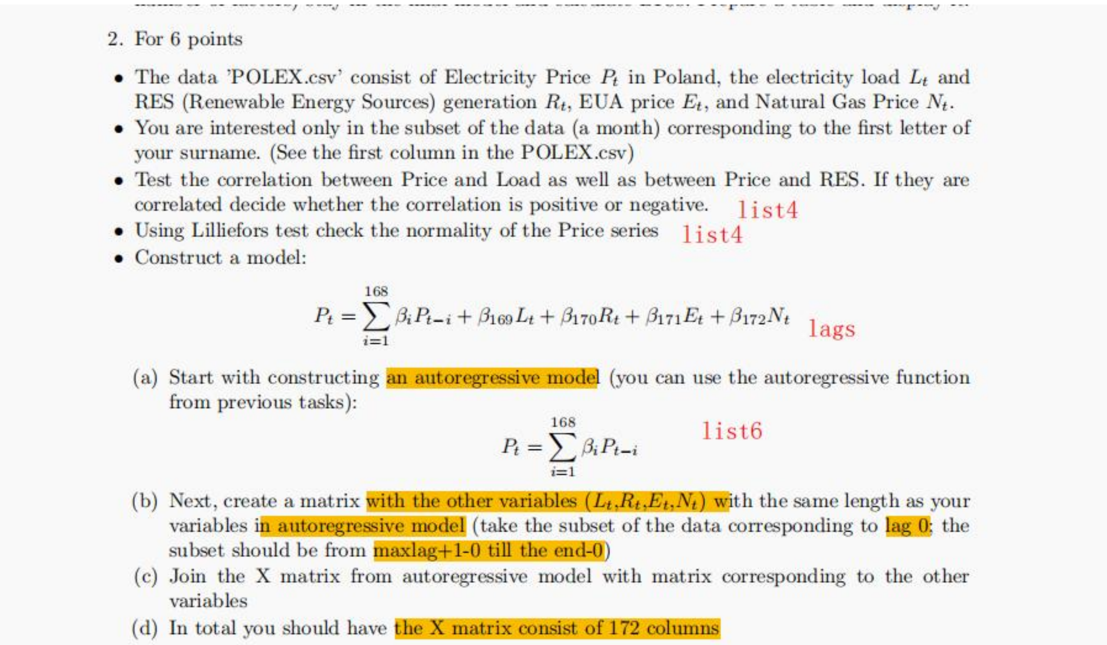
Task 5.3: Dimension reduction (6 points):

1. For 3 points

- The data 'POLEX.csv' consist of Electricity Price P_t in Poland
- You are interested only in the subset of the data (a month) corresponding to the first letter of your surname. (See the first column in the POLEX.csv)
- Construct a model:

$$P_t = \sum_{i=1}^{168} \beta_i P_{t-i}$$

- Estimate 10 factors from the matrix of independent variables
- Plot 3 first loadings
- Use criteria to assess optimal number of factors. Interpret the result (the conclusions can differ for both criteria)
- With OLS estimate the model consisting of optimal number of factors ($P_t = \sum_{i=1}^{\text{optimal}k} \beta_i F^i$)
- Estimate lasso for default grid of λ
- Select the optimal λ with AIC
- Select the optimal λ with CV with 5 folds
- Select the optimal λ with corrected version of CV with 5 folds
- For each of these 5 models (2xPCA + 3xLASSO) recognize how many variables (for PCA number of factors) stay in the final model and calculate BICs. Prepare a table and display it.



number of factors) stay in the final model and calculate BICs. Prepare a table and display it.

2. For 6 points

- The data 'POLEX.csv' consist of Electricity Price P_t in Poland, the electricity load L_t and RES (Renewable Energy Sources) generation R_t, EUA price E_t, and Natural Gas Price N_t.
- You are interested only in the subset of the data (a month) corresponding to the first letter of your surname. (See the first column in the POLEX.csv)
- Test the correlation between Price and Load as well as between Price and RES. If they are correlated decide whether the correlation is positive or negative. list4
- Using Lilliefors test check the normality of the Price series list4
- Construct a model:

$$P_t = \sum_{i=1}^{168} \beta_i P_{t-i} + \beta_{169} L_t + \beta_{170} R_t + \beta_{171} E_t + \beta_{172} N_t \quad \text{lags}$$

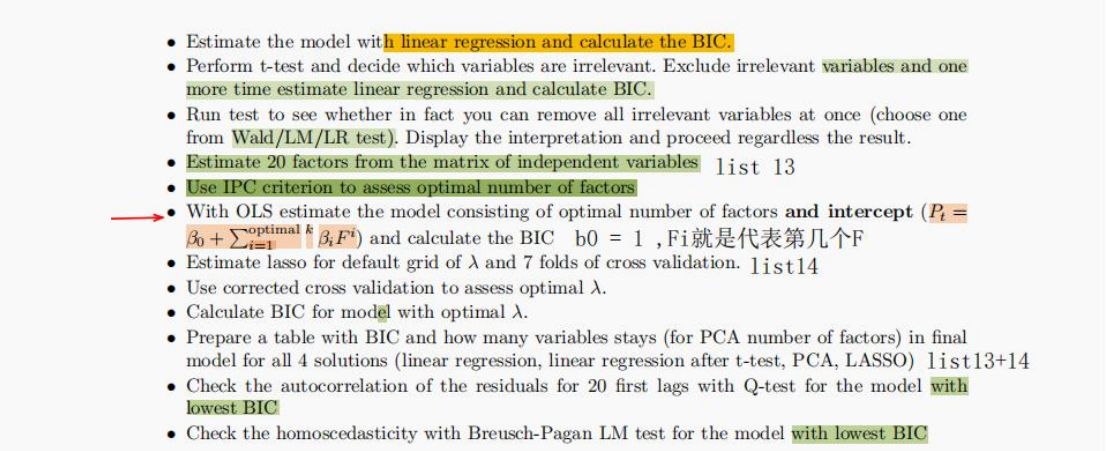
(a) Start with constructing an autoregressive model (you can use the autoregressive function from previous tasks):

$$P_t = \sum_{i=1}^{168} \beta_i P_{t-i} \quad \text{list6}$$

(b) Next, create a matrix with the other variables (L_t,R_t,E_t,N_t) with the same length as your variables in autoregressive model (take the subset of the data corresponding to lag 0: the subset should be from maxlag+1-0 till the end-0)

(c) Join the X matrix from autoregressive model with matrix corresponding to the other variables

(d) In total you should have the X matrix consist of 172 columns



- Estimate the model with linear regression and calculate the BIC.
- Perform t-test and decide which variables are irrelevant. Exclude irrelevant variables and one more time estimate linear regression and calculate BIC.
- Run test to see whether in fact you can remove all irrelevant variables at once (choose one from Wald/LM/LR test). Display the interpretation and proceed regardless the result.
- Estimate 20 factors from the matrix of independent variables list 13
- Use IPC criterion to assess optimal number of factors
- With OLS estimate the model consisting of optimal number of factors **and intercept** ($P_t = \beta_0 + \sum_{i=1}^{\text{optimal } k} \beta_i F^i$) and calculate the BIC b0 = 1 ,Fi就是代表第几个F
- Estimate lasso for default grid of λ and 7 folds of cross validation. list14
- Use corrected cross validation to assess optimal λ.
- Calculate BIC for model with optimal λ.
- Prepare a table with BIC and how many variables stays (for PCA number of factors) in final model for all 4 solutions (linear regression, linear regression after t-test, PCA, LASSO) list13+14
- Check the autocorrelation of the residuals for 20 first lags with Q-test for the model with lowest BIC
- Check the homoscedasticity with Breusch-Pagan LM test for the model with lowest BIC