Applied Statistics (ECS764P) - Lab 4

Fredrik Dahlqvist

8 Dec 2022

1 Theory

1. Consider the measure d on $\{1,2,3\} \times \{1,2,3\}$ defined by

$$\begin{array}{lll} d(1,1) = {}^1\!/_{10} & d(1,2) = {}^2\!/_{10} & d(1,3) = {}^1\!/_{10} \\ d(2,1) = {}^1\!/_{10} & d(2,2) = {}^1\!/_{10} & d(2,3) = {}^2\!/_{10} \\ d(3,1) = {}^1\!/_{10} & d(3,2) = 0 & d(3,3) = {}^1\!/_{10} \end{array}$$

(where I've written $d(x,y)$ for $d(\{(x,y)\})$ in order to keep things readable.)

(a) Is d a probability measure?

(b) Prove that d cannot be written as a product measure. *Hint: prove it by contradiction.*

(c) Compute the two marginals of d.

(d) Compute the covariance and correlation of d.

2 Practice

You can assume that `numpy`, `matplotlib`, `scipy` *and* `pandas-datareader` *are installed on the machine of the person who will run and mark your notebook. There is no need to force an install with the* `!` *command. For textual answers please use a markdown cell.*

1. You will first download the world GDP data from the World Bank using `pandas_datareader` (which we used in lab 2). The following code will download and plot the entire world GDP time series. Do NOT make any local copies of your data!

```
    from pandas_datareader import wb
    import matplotlib.pyplot as plt
    import numpy as np

    gdp_data = wb.download(indicator='NY.GDP.MKTP.CD', country='WLD',
start='1960', end='2021')
    time = np.arange(1960,2022)
    gdp = gdp_data.iloc[:,0].astype(float).to_numpy()
    # Data is returned in inverse chronological order, so reverse order
    gdp = np.flip(gdp)
    # Plot world GDP data against time
    plt.plot(time,gdp,label='US GDP')
    plt.legend()
    plt.show()

```

(you can ignore the warning about the code 'WLD'). You will try to estimate the long-term annual growth rate of the world using a regression.

- If the growth rate was a constant r, then the world's GDP would grow as

  $$GDP_k = GDP_0(1+r)^k$$

  where k is the number of years since 1960 and GDP_0 is the world's GDP in 1960. This is clearly not a linear relationship between time (k, in years) and GDP. However, we can get a linear relationship by applying a simple transformation $f(-)$ on both side of the equation. What is this transformation? (*Hint: we used this transformation in the context of MLE, it turns products into sums.*)
- Apply this transformation $f(-)$ to the GDP data, and perform a regression against the time variable. On the same plot, display your regression line, a scatter-plot of the (transformed) data points, and your R^2 value.
- Compute the residuals of your regression (i.e. the difference between the model and the observations), and print their mean and their standard deviation $\hat{\sigma}$. Perform a KS-test to determine whether we can reject the null hypothesis that the residuals are sampled from a normal distribution with mean 0 and standard deviation $\hat{\sigma}$. Take $\alpha = 99\%$.
- You will now apply the inverse of the transformation $f(-)$ to your linear model in order to get a non-linear model for the GDP. On the same plot, display your (non-linear) model and a scatter-plot of the (original) data points.
- What is the relationship between the slope of the regression and the long-term growth rate of the world GDP? Compute the long-term growth rate of the world GDP.
- What do you observe since approximately 2015?

2. Download the Dow Jones Industrial Average from Yahoo finance using the following code. Do NOT make any local copies of your data!

```
    import pandas_datareader.data as web
    import matplotlib.pyplot as plt

    data = web.DataReader('^DJI', 'yahoo', start='1995-01-01', end='2022-12-05')
    data = data.reset_index()
    dates = data["Date"]
    dow = data["Close"].to_numpy()
    plt.plot(dates,dow)
    plt.show()

```

You will study the autocorrelation of this time series.

- Instantiate an array `lags=[1,2,3,5,10,15,20,30]`. For each lag in `lags` compute and plot the history of the 60-day rolling autocorrelation of `dow`. Make your plot(s) legible.
- For each lag in `lags`, compute the average of the autocorrelation times series computed in the previous step and use these 8 values to plot the autocorrelation function (autocorrelation against lag). What do you observe? Does this plot suggest that the Dow Jones is a white noise process? If not, can you suggest a better model?
- Compute and plot the daily returns of the Dow Jones (you did this in lab 2), repeat the previous two steps and answer the same questions for this time series instead.