STA503: Math & Stat for MBA II

Final Exam - Fall 2021

Question 1. Today is your first day as a data scientist at a super fast growing e-commerce start-up. The morning meeting brought you an understanding that the company is losing a lot of money for running cloud-hosted website. The CEO and CTO are seriously considering its own physical server, but they prefer to do the migration after they finalize series D funding. The management would like to see if the overflow of website visitors are due to periodic promotions or organic expansion of the service. If organic, they would like to act now, otherwise the migration will be postponed.

Your data science team is given to build a projection model. The data science team's senior colleague tells you that after weeks of research, she came up with the ARMA(1,5) process for daily website visitors (y_t):

$$y_t = \phi_1 y_{t-1} + e_t$$

where $e_t = v_t + \theta_1 v_{t-1} + \theta_2 v_{t-2} + \theta_3 v_{t-3} + \theta_4 v_{t-4} + \theta_5 v_{t-5}$ and v_t is an innovation.

1) Discuss why it is important to test for non-stationarity and discuss the consequences of non-stationarity in regression analysis [5 marks]

2) Discuss the condition for this process to be covariance-stationary [5 marks]

3) Assuming the stationarity condition is satisfied, discuss how one can obtain consistent (not necessarily efficient) estimates of the ϕ parameters [5 marks]

4) Assume that all θs are statistically significant and positive. Draw potential ACF and PACF for e_t, and based on two graphes, can you suggest any alternative form of the model? [5 marks]

You, on the other hand, feel that the suggested ARMA model is odd by two facts. Considering the fact that usual online shopping malls become double busy on weekends that you learned from SIAI's AI Marketing course, you would like to carve out week-day/-end effects. Besides, to boost sales, in an attempt of window dressing for Series D funding, the company has been running Saturday Night Live 1+1 (SNL 1+1) promotion.

5) How would you modify the model? Provide specific details of the estimation steps. Why is this necessary? [5 marks]

6) How do you compare your modification to your senior colleagues? Provide your logic in both frequentist and Bayesian style arguments. Which one do you prefer for this company? [5 marks]

7) Assuming that the e-commerce has a fascinating data engineer team that gives you impeccable visitor data on hourly basis. Between frequentist and Bayesian, which option do you prefer? If you can't find a reasonable guess for SNL visitors, do you change your preference? [5 marks]

8) Data engineer team leader suggests you if collecting minute by minute data can be of any better use. What is your opinion, assuming that you have a 0.5% share of the company, so you care about cost saving. [5 marks]

Among one of the investors from the Series C round, one investor demands the management to run AI model for visitor projection, which can be a convincing signal for new investors that the e-commerce has AI experts. He does not like to use "traditional statistical models", such as AR(I)MA and seasonality decomposition, and asks the CEO to apply a "new and fancy" deep-learning model.

9) Without any data pre-processing, what do you expect the "AI model" will be? Can you guarantee that you can come up with re-usable model? [5 marks]

10) After your little demonstration in 9), the venture capitalist finally stepped back a bit, and asks what pre-processing you can propose. Prove your worthiness for both ARMA and seasonality cases. [5 marks]

Question 2. Consider the following linear regression model

$$y_t = \beta_0 + \rho y_{t-1} + \beta_1 z_t + \beta_2 z_{t-1} + \beta_3 z_{t-2} + \epsilon_t \text{ with } |\rho| < 1$$

with ϵ_t a covariance stationary mean zero process. You may assume there is no perfect multi-collinearity in the regressors. As a marketing team data scientist, you propose a prediction model for your client's advertisement spending (y_t), as a function of your client's website visits (z_t). Both y_t and z_t are monthly basis.

1) Discuss what it means to say that ϵ_t is a covariance stationary mean zero process. [5 marks]

2) Indicate, in terms of our parameters, what the marginal effect of z on y are, both short and long-run. What do these effects tell us? [5 marks]

3) Why you would not want to impose $\mathbb{E}(\epsilon_t|X) = 0$ in this setting. [5 marks]

4) Do you agree that it really is a covariance stationary process? If not, what are the necessary data pre-processing steps can you propose? If you agree, why do you think data pre-processing is unnecessary? [5 marks]

5) Your client's marketing team periodically change the members. Given that, your data science team leader believes lagged variables are redundant. His proposal of a new model is to set $\rho = 0$, that is, consider

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 z_{t-1} + \beta_3 z_{t-2} + \eta_t$$

Do you agree with his model? If not, propose any other model that incorporates your client specific situation. [5 marks]

You also have a new team leader from traditional marketing business whose experience in online marketing is as good as Australopithecus. To cover up his ignorance, the boss hired an ultra expensive consulting company, but the consultants are just as bad as the Australopithecus deserves. The consultants, with no statistical but IT business training, argue that all data are needed for 99.9% accuracy of the proposed AI model, which is nothing more than a deep-learning coding library.

6) How would you convince your boss that blind deep-learning will fail? [5 marks]

7) Let's assume that the Australopithescus accepts your logic not by a divine miracle or your robustness but by budget pressure from the finance team. Now the boss wants to report the projection as soon as possible. He still does not buy your idea that website visitors are key factors. He demands you to come up with a model only with past advertisement spending. Below is your compromise.

$$y_t = \beta_0 + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \delta_t \text{ with } |\rho_i| < 1 \text{ for all } i$$

What are the merits and demerits of the further lags? If above model somehow turned out to be a long-term sustainable model, when $\rho_1 = 2, \rho_2 = -1$, what will be the cycle of the process? [5 marks]

8) Your boss does not like to continue with the cycle in 7), because it means your client will go in downturn for the next a year, which results in a large haircut in your boss's next year bonus. He wonders if you can change the cycle by adding further lagged variables. Based on what you have in 7), what will be the results with further lags? [2 marks]

9) Given your argument, your boss gave up hiding ignorance and decides to add back the visitors. Let us consider the original model again with $\rho \neq 0$. In the middle of research in all possibilities of ARMA, you found that ϵ_t follows an MA(2) process, i.e., $\epsilon_t = v_t + \theta_1 v_{t-1} + \theta_2 v_{t-2}$ where v_t is white noise $(0, \sigma_v^2)$ independent of $y_{t-1}, y_{t-2}, ...$; v_t is also independent of z_s for all s.

   Describe how you would obtain consistent parameter estimates for the parameters θ, and (β, ρ) [8 marks]

10) What happens if ϵ_t follows an AR(1) process? In this situation, what would you like to propose? [5 marks]