NATIONAL UNIVERSITY OF SINGAPORE

BT1101 – INTRODUCTION TO BUSINESS ANALYTICS

Semester 1: AY2020/21

25 November 2020

Time Allowed: 150 min

INSTRUCTIONS TO STUDENTS

You are free to allocate 150 mins on 20 Multiple Choice Questions and 3 Comprehensive Questions. You can navigate between all 23 questions.

2. Please answer ALL questions.
3. Please submit all your answers in Examplify **when exam time is up and before 7:40 PM**.
4. This is **NON-SECURE BLOCK INTERNET** assessment. Questions are **randomly sequenced**.
5. Make your best guess and do not spend too much time on a single question.
6. For Comprehensive Questions, the description or context of each question is followed by a series of subquestions labeled with (a), (b), (c), etc. Your submitted answers should be clearly labeled. At the bottom of each question, the Rmarkdown code for the question itself is provided. **Please copy and paste the Rmarkdown code for each question to your Rmarkdown template since only Examplify submission will be graded.**
7. You should program and type all your answer in **Final Exam Rmarkdown Template file** in the order of questions generated.
8. **IMPORTANT:** you need to **include your entire Rmarkdown code** (r-chunks and your answers) for each question in Examplify before your submission in Examplify. Simply copy and paste your answer from the Rmarkdown template to Examplify.
9. **IMPORTANT:** Please s**ubmit both your RMD and HTML file in the folder "Final Exam Submission"** in Luminus after your submission in Examplify and Internet reconnection.
10. Once you have completed the assessment, we suggest you to leave plenty of time for Examplify submission:
    - Click on the "Exam Controls" button and choose "Submit Exam".
    - Check off “I confirm that I have completed my exam” and click on “Submit Exam“ to upload your answers to the server.
    - You will see a green confirmation window on your screen when the upload is successful.
    - If you do not see a green window, please disconnect and reconnect your WIFI and try again.
    - Please be reminded that it is your responsibility to ensure that you have uploaded your answers to the Software.
11. Please switch off your personal devices with communication features at all times except to call the examiner

or technical team or to contact them on MS team.

12. You will be liable for disciplinary action which may result in expulsion from the University if you are found to have contravened any of the clauses below:

- Violation of the NUS Code of Student Conduct (in particular the part on Academic, Professional and Personal Integrity), NUS IT Acceptable Use Policy or NUS Examination rules.
- Possession of unauthorized materials/electronic devices.
- Bringing your mobile phone or any storage/communication device with you to the washroom.
- Unauthorized communication e.g. with another student.
- Photography or videography during the exam.
- Reproduction of any exam materials after the exam.
- Plagiarism, giving or receiving unauthorised assistance in academic work, or other forms of academic dishonesty.

Thank You!

STUDENT NO: _____

This portion is for examiner's use only

Section B	**Marks**	**Remarks**
Question 1		
Question 2		
Question 3		
Total		

Question #: 1

I agree that I have read and will abide by the **Code of Student Conduct** (in particular the part on Academic, Professional and Personal Integrity), as well as items B and C below.

(A) **I am aware of, and will abide by the NUS Code of Student Conduct (in particular the part on Academic, Professional and Personal Integrity as shown below) when attempting this assessment.**

- Academic, Professional and Personal Integrity
    1. The University is committed to nurturing an environment conducive for the exchange of ideas, advancement of knowledge and intellectual development. Academic honesty and integrity are essential conditions for the pursuit and acquisition of knowledge, and the University expects each student to maintain and uphold the highest standards of integrity and academic honesty at all times.
    2. The University takes a strict view of cheating in any form, deceptive fabrication, plagiarism and violation of intellectual property and copyright laws. Any student who is found to have engaged in such misconduct will be subject to disciplinary action by the University.
    3. It is important to note that all students share the responsibility of protecting the academic standards and reputation of the University. This responsibility can extend beyond each student's own conduct, and can include reporting incidents of suspected academic dishonesty through the appropriate channels. Students who have reasonable grounds to suspect academic dishonesty should raise their concerns directly to the relevant Head of Department, Dean of Faculty, Registrar, Vice Provost or Provost.

(B) **I have read and understood the rules of the assessments as stated below.**

1. Students should attempt the assessments on their own. There should be no discussions or communications, via face to face or communication devices, with any other person during the assessment.
2. Students should not reproduce any assessment materials, e.g. by photography, videography, screenshots, or copying down of questions, etc.

(C) **I understand that by breaching any of the rules above, I would have committed offences under clause 3(l) of the NUS Statute 6, Discipline with Respect to Students which is punishable with disciplinary action under clause 10 or clause 11 of the said statute.**
3) Any student who is alleged to have committed or attempted to commit, or caused or attempted to cause any other person to commit any of the following offences, may be subject to disciplinary proceedings:
i) plagiarism, giving or receiving unauthorised assistance in academic work, or other forms of academic dishonesty.

Type your name to declare that you have read and will abide by the NUS Code of Student Conduct (in particular, (a) Academic, Professional and Personal Integrity), (b) and (c).

Question #: 22

Instruction:
1. Copy and paste the Question Rmarkdown Code (at the bottom) into your Rmarkdown Template for the Final Exam.
2. Make sure you include all your answer (r-chunk and text answers) from your Rmarkdown file back to the Essay Answer section in Examplify.
3. Please submit both your Rmarkdown and HTML files in Luminus folder "Final Exam Submission" before 7:40 PM.

Context
University GPA (total 15 marks)
- Data set: `gpa2` in `wooldridge` public data sets.

```{r load-gpa2}
# load the data set, make sure you already loaded `wooldridge` package
data(gpa2)
```

This data set is from a midsize research university. It has 4137 observations on 12 variables:

- `sat`: combined SAT score (includes verbal, writing and maths score)
- `tothrs`: total hours through fall semester
- `colgpa`: GPA after fall semester
- `athlete`: =1 if athlete
- `verbmath`: verbal and math SAT score
- `hsize`: size of high school graduation class, 100s
- `hsrank`: rank in high school graduation class (where rank 1 is top in the class)
- `hsperc`: high school percentile, from top (i.e. a value of "10" means "top 10 percent in high school")
- `female`: =1 if female
- `white`: =1 if white
- `black`: =1 if black
- `hsizesq`: hsize^2

(a) This dataset comprises 4 demographic variables (`athlete`, `female`, `white`, `black`) on students.
i) Create a table to display the frequency of students in the dataset for each category defined by the combination of all 4 variables. You can use a normal table or a pivot table. You may exclude combination(s) with no occurrence in the table. (1 mark)
ii) Based on the table in (ai), what is the difference in number of black male athlete students to non-black male athlete students? (1 mark)

(b) There are a few variables that measure the performance of students, namely `sat`, `colgpa`, `verbmath`, `hsrank` and `hsperc`. You may treat these as continuous random variables.
i) The university is interested to know if there is any linear relationship between `colgpa` and high school performance (`sat`, `verbmath`, `hsrank` and `hsperc`). Check this visually as well as with the appropriate statistical measure(s). Interpret your results. (2 marks)
ensure all graphs are clearly labeled with the appropriate titles and axes names.
ii) Compute and interpret the 99% prediction interval for `colgpa`. (2 marks)

(c) Set up and test the following hypotheses:
i) Is the mean `colgpa` for male athlete students different from male non-athlete students? (1.5 marks)
ii) Is the proportion of students with a `colgpa` of more than 3.5, less than 12%? Use alpha=0.01 (2.5 marks)

d) The university admin office divides the students into 4 categories, along these two demographic variables - `athlete` and `white`.

i) Compare using a graph, the standard deviations of `colgpa` across the different categories of students. Describe your observation from the graph. (2 marks)
ensure all graphs are clearly labeled with the appropriate titles and axes names.
ii) With the sample data available, what can you conclude about the statement that “the variation of `colgpa` is not the same across the four categories of students”? (3 marks)

Question Code To Be Pasted in Rmarkdown:

```
## Question BT1101- University GPA (total 15 marks)

- Data set: `gpa2` in `wooldridge` public data sets.

```{r load-gpa2}
# load the data set, make sure you already loaded `wooldridge` package
data(gpa2)
```

This data set is from a midsize research university. It has 4137 observations on 12 variables:

- `sat`: combined SAT score (includes verbal, writing and maths score)
- `tothrs`: total hours through fall semest
- `colgpa`: GPA after fall semester
- `athlete`: =1 if athlete
- `verbmath`: verbal and math SAT score
- `hsize`: size of high school graduation class, 100s
- `hsrank`: rank in high school graduation class (where rank 1 is top in the class)
- `hsperc`: high school percentile, from top (i.e. a value of "10" means "top 10 percent in high school")
- `female`: =1 if female; =0 if male
- `white`: =1 if white
- `black`: =1 if black
- `hsizesq`: hsize^2

(a) This dataset comprises 4 demographic variables (`athlete`, `female`, `white`, `black`) on students.

- **i) Create a table to display the frequency of students in the dataset for each category defined by the combination of all 4 variables. You can use a normal table or a pivot table. You may exclude combination(s) with no occurrence in the table.** (1 mark)

- **ii) Based on the table in (ai), what is the difference in number of black male athlete students to non-black male athlete students?** (1 mark)

Type your answer here

(b) There are a few variables that measure the performance of students, namely `sat`, `colgpa`, `verbmath`, `hsrank` and `hsperc`. You may treat these as continuous random variables.

- **i) The university is interested to know if there is any linear relationship between `colgpa` and high school performance (`sat`, `verbmath`, `hsrank` and `hsperc`). Check this visually as well as with the appropriate statistical measure(s). Interpret your results.** (2 marks)
(ensure all graphs are clearly labeled with the appropriate titles and axes names.)

- **ii) Compute and interpret the 99% prediction interval for `colgpa`.** (2 marks)

Type your answer here

(c) Set up and test the following hypotheses:

- **i) Is the mean `colgpa` for male athlete students different from male non-athlete students?** (1.5 marks)

- **ii) Is the proportion of students with a `colgpa` of more than 3.5, less than 12%? Use alpha=0.01** (2.5 marks)

Type your answer here

d) The university admin office divides the students into 4 categories, along these two demographic variables - `athlete` and `white`.

- **i) Compare using a graph, the standard deviations of `colgpa` across the different categories of students. Describe your observation from the graph.** (2 marks)
(ensure all graphs are clearly labeled with the appropriate titles and axes names.)

- **ii) With the sample data available, what can you conclude about the statement that “the variation of `colgpa` is not the same across the four categories of students”?** (3 marks)
```

Question #: 23

Instruction:
1. Copy and paste the Question Rmarkdown Code (at the bottom) into your Rmarkdown Template for the Final Exam.
2. Make sure you include all your answer (r-chunk and text answers) from your Rmarkdown file back to the Essay Answer section in Examplify.
3. Please submit both your Rmarkdown and HTML files in Luminus folder "Final Exam Submission" before 7:40 PM.

Context

Portfolio Management (total 15 marks)
Consider you are a portfolio manager in charge of a simple portfolio which consists of two public traded stocks in Singapore Exchange (SGX) market, 5Xnergy (SGX: 5X) and EverGreen Tech (SGX: EG). The stocks are traded in minimum unit of one share. Given the current and predicted stock prices, the portfolio manager who starts with zero holding, decides the holding positions of the stocks in portfolio, i.e. the number of shares, at the beginning of a financial year and holds the portfolio for a year. Below is the information about the two stocks:

Stock (per share)	Current Price (SGD)	Predicted Price in 1yr (SGD)	Risk (SGD)
5X	15.6	19.2	0.37
EG	3.5	21.9	18.21

Risk of a stock is the standard deviation of the predicted price in one year. Assume that the total risk of the portfolio is a linear combination of risk of the stocks in the portfolio, weighted by the positions, i.e. total risk of a portfolio consisting of a #shares of 5X and b #shares of EG is 0.37a + 18.21b. The total risk of the portfolio should not exceed 50,000 SGD. The portfolio manager is endowed with an investment budget of one million SGD and tries to maximize the total return of the portfolio.

(a) Write down the decision variables, the objective function, and ALL relevant constraints that apply for this optimization problem in a table formulation. Do NOT solve the problem yet. (5 marks)

(b) Solve your formulated optimization problem in R. What are the optimal holdings of the two stocks in the portfolio? What is the optimal total return of the portfolio? (3 marks)

(c) Predicted price in one year for each stock is naturally random from the current perspective of the portfolio manager. Suppose that predicted price in one year of each stock follows a log-normal distribution with mean mu_i and sigma_i where i = 5X, EG, provided in the table below.

Stock	μ	σ
5X	19	0.3
EG	28	20

Use Monte Carlo method to simulate 100 predicted prices in one year for both stocks and answer the following questions. (2 marks)

- What is the simulated standard deviation of predicted price for each stock, i.e. risk of each stock? (1 mark)
- What is the *average* optimal total return of the portfolio? (3 marks)
- What is the probability that the optimal total return of the portfolio is less than 250,000 SGD? (1 mark)

Question Code To Be Pasted in Rmarkdown:

Question: Portfolio Mangement (total 15 marks)
Consider you are a portfolio manager in charge of a simple portfolio which consists of two public traded stocks in Singapore Exchange (SGX) market, 5Xnergy (SGX: 5X) and EverGreen Tech (SGX: EG). The stocks are traded in minimum unit of one share. Given the current and predicted stock prices, the portfolio manager who starts with zero holding, decides the holding positions of the stocks in portfolio, i.e. the number of shares, at the beginning of a financial year and holds the portfolio for a year. Below is the information about the two stocks:

Stock (per share)	Current Price (SGD)	Predicted Price in 1yr (SGD)	Risk (SGD)
5X	15.6	19.2	0.37
EG	3.5	21.9	18.21

Risk of a stock is the standard deviation of the predicted price in one year. Assume that the total risk of the portfolio is a linear combination of risk of the stocks in the portfolio, weighted by the positions, i.e. total risk of a portfolio consisting of a \#shares of 5X and b \#shares of EG is $0.37a + 18.21b$. The total risk of the portfolio should not exceed 50,000 SGD. The portfolio manager is endowed with an investment budget of one million SGD and tries to maximize the total return of the portfolio.

(a) Write down the decision variables, the objective function, and ALL relevant constraints that apply for this optimization problem in a table formulation. Do NOT solve the problem yet. (5 marks)

Type your answer here

(b) Solve your formulated optimization problem in R. What are the optimal holdings of the two stocks in the portfolio? What is the optimal total return of the portfolio? (3 marks)

Type your answer here

(c) Predicted price in one year for each stock is a naturally random variable from the current perspective of the portfolio manager. Suppose that predicted price in one year of each stock follows a log-normal distribution with mean μ_i and σ_i where $i = 5X, EG$, provided in the table below.^[Recall that `rlnorm(n, meanlog, sdlog)` in R generates random numbers from a log-normally distributed variable $X \sim \text{LogNormal}(\mu_X, \sigma_X^2)$ where `meanlog` $=\ln\left(\frac{\mu_X^2}{\sqrt{\mu_X^2 + \sigma_X^2}}\right)$ and `sdlog` $=\sqrt{\ln\left(1 + \frac{\sigma^2_X}{\mu_X^2}\right)}$. μ_X and σ_X^2 are the mean and variance of the log-normally distributed X, respectively.]

Stock	μ	σ
5X	19	0.3
EG	28	20

Use Monte Carlo method to simulate 100 predicted prices in one year for both stocks and answer the following questions. (2 marks)

- **What is the simulated standard deviation of predicted price for each stock, i.e. risk of each stock?** (1 mark)

- **What is the *average* optimal total return of the portfolio?** (3 marks)

- **What is the probability that the optimal total return of the portfolio is less than 250,000 SGD?** (1 mark)

Type your answer here

```{r, echo=TRUE}
set.seed(1)
n_sample = 100
# ...
```

Question #: 24

Instruction:
1. Copy and paste the Question Rmarkdown Code (at the bottom) into your Rmarkdown Template for the Final Exam.
2. Make sure you include all your answer (r-chunk and text answers) from your Rmarkdown file back to the Essay Answer section in Examplify.
3. Please submit both your Rmarkdown and HTML files in Luminus folder "Final Exam Submission" before 7:40 PM.

Context

Traffic Laws (total 15 marks)
- Data set: `*traffic2*` in *wooldridge*.

```{r loaddta}
# load the data set, make sure you already load `wooldridge` package
data(traffic2)
```

This data set contains 108 monthly time-series observations with 48 variables on state-wide traffic accidents. For this question, the relevant variables are the following:

- `year`: 1981 to 1989
- `totacc`: total number of accidents
- `t`: time trend
- `spdlaw`: = 1 after 65 mph law in effect
- `beltlaw` = 1 after seatbelt law in effect

(a) Traffic regulation policymaker is concerned if laws on speeding and wearing seatbelt have effect on the number of road accidents. Using the following four variables `totacc`, `t`, `spdlaw` and `beltlaw`, run a linear regression model to examine the relationship. Report the regression output and write out the *fitted line*. (4 marks)

(b) Interpret the coefficient estimators before `spdlaw` and `beltlaw`, respectively. (2 marks)
Assuming the model is valid, please explain (by proposing possible theory) and make sense of the sign (direction of the effect) of coefficient estimators before `spdlaw` and `betlaw`. (2 marks)

(c) From your regression output alone without checking assumptions, why do you think we need to include the time trend `t` in the regression? (1 mark)

(d) Now let's single out the time series variable `totacc`, the monthly total number of accidents between 1981 and 1989.

```{r ts_obj}
# define `totacc` as a monthly time series object
totacc = ts(traffic2$totacc, frequency = 12, start = 1981)
```
Plot the time series `totacc` and describe if the time series `totacc` shows any trend, seasonality or cyclicality. (3 marks)

(e) Following the steps of best practice, identify an ARIMA model for univariate time series `totacc`. Provide justification with your proposed ARIMA model. Ignore the seasonality if any. (3 marks)

Question Code To Be Pasted in Rmarkdown:

Question: Traffic Laws (total 15 marks)
- Data set: `traffic2` in `wooldridge` public data sets.

```{r loaddta}
# load the data set, make sure you already load `wooldridge` package
data(traffic2)
```

This data set contains 108 monthly time-series observations with 48 variables on state-wide traffic accidents. For this question, the relevant variables are the following:

- `year`: 1981 to 1989
- `totacc`: total number of accidents
- `t`: time trend
- `spdlaw`: = 1 after 65 mph law in effect
- `beltlaw` = 1 after seatbelt law in effect

(a) Traffic regulation policymaker is concerned if laws on speeding and wearing seatbelt have effect on the number of road accidents. Using the following four variables `totacc`, `t`, `spdlaw` and `beltlaw`, run a linear regression model to examine the relationship. Report the regression output and write out the *fitted line*. (4 marks)

Type your answer here

(b) Interpret the coefficient estimators before `spdlaw` and `beltlaw`, respectively. (2 marks)
Assuming the model is valid, please explain (by proposing possible theory) and make sense of the sign (direction of the effect) of coefficient estimators before `spdlaw` and `betlaw`. (2 marks)

Type your answer here

(c) From your regression output alone without checking assumptions, why do you think we need to include the time trend `t` in the regression? (1 mark)
Type your answer here

(d) Now let's single out the time series variable `totacc`, the monthly total number of accidents between 1981 and 1989.
```{r ts_obj}
# define `totacc` as a monthly time series object
totacc = ts(traffic2$totacc, frequency = 12, start = 1981)
```
- **Plot the time series `totacc` and describe if the time series `totacc` shows any trend, seasonality or cyclicality. (3 marks)**

Type your answer here

(e) Following the steps of best practice, identify an ARIMA model for univariate time series `totacc`. Provide justification with your proposed ARIMA model. Ignore the seasonality if any. (3 marks)

Type your answer here