**MS5318 Final Exam**

2020-2021 Semester B

**Please read the following guidelines for the final exam, and sign the honor pledge.**

1. This exam is open-book (limited to our lecture notes).
2. You will need to use Excel and R to solve the problems.
3. Distribution of this exam paper (including the dataset) to anybody is prohibited.
4. You are **required to join the Zoom meeting** during the final exam **with your camera open**. The zoom meeting is posted on Canvas.
5. Final exam duration is 3 hours from 6:30 pm to 9:30 pm. Time constraint will be strictly enforced. Hence, I encourage you to submit early. Note that you can submit multiple times through Canvas, and only the last submission counts. Late submission:**10% of your scores will be deducted**. **The exam submission link will be disabled at 9:40pm. Submission after that will not be accepted.**
6. Your solutions (including all R codes) should be compiled in this word file.
7. If you cannot login Canvas, send your exam answers through email: [zhankun.sun@cityu.edu.hk](mailto:zhankun.sun@cityu.edu.hk) by the due time. Late submissions will be dealt with in the same manner as described in item 4.
8. No collaboration or online communications is allowed.
9. In case of emergency, contact me at +852 3442 8650.
10. The departmental hotline is +852 3442 8325.

|  |
| --- |
| **CityU Honor Pledge:** I affirm that I will not give or receive any unauthorized help on this exam, and that all work will be my own. A direct result of any violation of the honor pledge is failing this course. |
| **Signature by tying in your full name:** |

|  |
| --- |
| Question 1. (10pts) data file: poets.xlsx  Do poets die young? According to William Butler Yeats, “She is the Gaelic muse, for she gives inspiration to those she persecutes. The Gaelic poets die young, for she is restless, and will not let them remain long on earth.” One study designed to investigate this issue examined the age at death for writers from different cultures and genders. Three categories of writers examined were novelists, poets, and nonﬁction writers. The ages at death for female writers in these categories from North America are given in poets.xls. Most of the writers are from the United States, but Canadian and Mexican writers are also included  Does the mean age at death differ among the three groups? Run the appropriate procedure and summarize the findings. Use a significance level α=0.05. In your submission, include the Excel output tables NOT the original data. |
| Answer: |
| **Question 2. (10pts)**  A particular paperback mystery book is published with a choice of three different pictures on the cover: a **photograph** of the actor playing the main character in the movie version of the book, a **drawing** of the mansion where the story in the book takes place, or an **embossed graphic** of the murder weapon. A certain bookstore keeps copies of this book with each of the pictures on the cover on its racks. To test the hypothesis that sales of this book are equally divided among the three choices, a simple random sample of 120 purchases of this book is obtained. The numbers of photograph, drawing, and embossed graphic are respectively 31, 47, and 42.  Run the appropriate procedure to test the null hypothesis that sales of this book are equally divided among the three choices. Write down the null and alternative hypothesis. Report the results and the conclusion of your hypothesis tests below. Use . In your submission, include the Excel output tables and name it Q2.xlsx. |
| **Answer:** |

|  |
| --- |
| Question 3. (40pts) data file: leases.csv  This data table gives annual costs of 223 commercial leases. All of these leases provide office space in a Midwestern city in the United States. For the response, use Lease.Cost (the cost of the lease). As explanatory variables, use Square.Feet (the number of square feet) and Age (the age in years of the property in which the office space is located).  Provide necessary R output and your answers to the box below each question. All R codes go to the last box. |
| (a)(5pts) Examine scatterplots of the response versus the two explanatory variables as well as the scatterplot between the explanatory variables. Do you think it appropriate to fit a multiple linear regression model? (Attach the scatterplot here) |
| Answer: |
| (b)(15pts) Fit the indicated multiple regression and show a summary of the estimated model. Remember to verify the modeling assumptions for linear regressions and report your conclusions. Does this estimated model explain statistically significant variation in the cost of the leases? |
| Answer: |
| (c)(10pts) Fit a regression model of **Lease.Cost** on **Age**. Compare the marginal coefficient for **Age** to the partial coefficient obtained in part (b). Explain why these are different in magnitude and sign. |
| Answer: |
| (d)(10pts) Some of these leases cover space in the downtown area, whereas others are located in the suburbs. The variable **Location** identifies these two categories. Fit a multiple regression model including Square.Feet, Age, Location, the interaction between Square.Feet and Location, and the interaction between Age and Location. (Attach the model summary below.)  Interpret the estimated coefficient for the interaction between Square.Feet and Location of the model. What does the estimated value mean? |
| Answer: |
| # R Codes for Question 3 |
| **Question 4 (40 pts) data file: admission.csv, admission\_testing.csv**  An emergency department (ED) physician is interested in what factors can best predict whether a patient will be admitted to the hospital or discharge home after the patient’s treatment in the ED. The predictors include patient **age** (at the time of ED visit), **sex** (male or female), **triage level** (a score indicating the patient’s urgency level), whether the patient took an ambulance to the ED or not, whether the patient was treated in the fast-track area of the ED or not, and finally the patient’s address. The response variable, **admit**, is a binary variable (admit = 0 means the patient is discharged home, admit = 1 means the patient got admitted to the hospital). The predictors in our data are   * **age**: the patient’s age at the time of the visit (numerical variable) * **sex**: F represents female; M represents male (categorical variable) * **CTAS**: the triage level from 1 to 5, with 1 indicates most urgent and 5 indicates least urgent (categorical variable with five levels) * **EMS**:whether the patient took ambulance or not. Y means yes, and N means no. * **FT**: whether the patient is treated in the fast-track area or not. Y means yes, and N means no. * **address**: the region in the study city (categorical variable with five levels)   Provide necessary R output and your answers to the box below each question. All R codes go to the last box. |
| (a)(5pts) Run a logistic regression to predict the probability of admission to the hospital using all the given predictors excluding their interactions and report the regression results. (use the dataset **admission.csv**)  (5pts) Explain the effects of age on the admission probability based on the model outputs. |
| Answer: |
| (b)(5pts) Note that to capture the nonlinear effect of age on the probability of admission, we divide age into five groups: 0<= age <18 years, 18<= age <40 years, 40<= age <55 years, 55<= age <70 years, and age>=70 years. Use the age group instead of age to re-run the model and report the regression results. (use the dataset **admission.csv**)  (5pts) Explain the effects of age group on the admission probability based on the model outputs. |
| Answer: |
| (c)(5pts) Using the model in part (b), predict the admission probability of the 410th patient in the testing dataset **admission\_testing.csv.** |
| Answer: |
| (d)(10pts) The logistic regression can predict the admission probability to the hospital given the predictors for each patient. For a given threshold **T**, if the predicted probability is greater than **T**, we say the predicted admission is 1 (admitted to the hospital); otherwise, 0 (discharged home).  We next define two concepts:  **Sensitivity** is also called the true positive rate or probability of detection. It measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of admitted patients who are correctly predicted as being admitted).  **Specificity** is also called the true negative rate, which measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of discharged patients who are correctly identified as being discharged).  Let the threshold be . Use the regression model developed in (b) to compute the sensitivity and specificity using the testing dataset **admission\_testing.csv**. Report your results below. |
| Answer: |
| (e)(5pts) Repeat part (d) using the model developed in (a) and report the sensitivity and specificity. |
| Answer: |
| # R Codes for Question 4 |