**MS5318 Final Exam**

2021-2022 Semester B

**Please read the following guidelines for the final exam:**

1. This exam is open-book and open-notes.
2. You are **required to join the Zoom meeting** during the final exam **with your camera open**. The zoom meeting is posted on Canvas.
3. Final exam duration is 3 hours from 6:30pm to 9:30pm. 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 9:45pm. Submission after that will not be accepted.**
4. Your solutions (including all R codes) should be compiled in this word file.
5. 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.
6. No collaboration or online communications is allowed.
7. Sign the honor pledge; otherwise, 10 points will be deducted from your final score.
8. In case of emergency, contact me at +852 3442 8650.
9. The departmental hotline is +852 3442 8585.

|  |
| --- |
| **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. True/False (10pts)  Mark each statement True or False. If you believe that a statement is false, briefly explain why you think it is false. |
| (a)(5pts) In order to calculate the VIF for an explanatory variable, we need to use the values of the response variable. |
| Answer: |
| (b)(5pts) A correlation matrix summarizes the same information in the data as is given in a scatterplot matrix. |
| Answer: |

|  |
| --- |
| Question 2. (45pts) data file: emerald\_diamonds.xls  This data table of 144 diamonds includes the Price (in dollars), the Weight (in carats), and the clarity grade of the diamonds. The diamonds have clarity grade either VS1 or VVS1. VVS1 diamonds are nearly flawless; VS1 diamonds have more visible (but still small) flaws.  Provide necessary R output and your answers to the box below each question. All R codes go to the last box. |
| (a)(10pts) Would it be appropriate to use a two-sample t-test to compare the average prices of VS1 and VVS1 diamonds, or is this relationship confounded by the weights of the diamonds? Provide evidence to support your conclusion. |
| Answer: |
| (b)(10pts) Perform the two-sample t-test to compare the prices of the two grades of diamonds. Summarize this analysis as if there are no confounding variables. |
| Answer: |
| (c)(10pts) Include Weight into the comparison. Determine if the interaction should be kept in the model (show your model output and justify why you want (or not) to keep it). With the final model, compare the prices of the two types of diamonds. |
| Answer: |
| (d)(15pts) Compare the results from parts (b) and (c). What can you conclude about the cost of diamonds of these two grades? |
| Answer: |
| # R Codes for Question 2 |
| **Question 3 (45 pts) data file: germancredit.csv**  The German credit data set contains attributes and outcomes on 1000 loan applications. It has served as an important data set for several credit-scoring algorithms. The data set contains the following variables:   * Default: a binary variable indicating bad/good loans. Default=1 indicates a bad loan, which means the loan did not pay as intended. * duration: duration of the loan * amount: amount of the loan * installment: installment of the loan * age: age of the borrower * history: credit history * purpose: purpose of the loan * foreign: foreign/german * rent: FALSE/TRUE   Split the data: use the first 500 observations as the **training dataset** and the rest of data as the **testing dataset**.  Provide necessary R output and your answers to the box below each question. All R codes go to the last box. |
| (a)(10pts) Fit a regression model **using the training dataset** with all the given variables (without interactions) to predict the Default probability. (Attach the regression outputs here) |
| Answer: |
| (b)(10pts) Which level of **purpose** has the largest predicted default probability, given that all other variables are fixed? Which level of **history** has the largest predicted default probability, given that all other variables are fixed? |
| Answer: |
| (c)(5pts) Using the model in part (a), predict the Default probability of the 800th case of the German credit data set. |
| Answer: |
| (d)(10pts) If the predicted Default probability is greater than or equal to 0.5, the loan will be classified as a bad loan: predicted Default=1. Otherwise, the loan will be classified as a good loan: predicted Default=0. With the model in part (a), we can classify each loan in the **testing dataset** into one of the two categories (predicted Default = 1 or 0). Using such a classification method, how many observations in the testing set are incorrectly classified? |
| Answer: |
| (e)(10pts) If the predicted Default probability is greater than or equal to a threshold **T**, the loan will be classified as a bad loan, i.e., PredictedDefault = 1. Otherwise, the loan will be classified as a good loan, i.e., PredictedDefault = 0. With the model in part (a), we can classify each loan in the **testing set** into one of the two classes (PredictedDefault = 1 or 0).  **Sensitivity,** also called the true positive rate or probability of detection,measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of actual bad loans in our data that are correctly predicted as bad loans by our model).  **Specificity,** also called the true negative rate, measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of good loans that are correctly identified as good loans by our model).  Let use R to compute the sensitivity and specificity of the model in (a) on the **testing set**.  Hint 1: To compute the sensitivity, you may first want to know the number of bad loans (say ). Then you want to compute the number of bad loans that are predicted as bad loan (say x). The sensitivity will be x/y. Similarly you can compute the specificity. Possible useful R functions: cbind() table() nrow()  Hint 2: Show your steps to get partial points |
| Answer: |
| # R Codes for Question 3 |