**MS5318: Final Exam Arrangement**

* This exam is open-book and open-notes. You have to **bring a computer with camera and use R** to solve some of the questions.
* You are **required to join the Zoom meeting** during the final exam **with your camera open**. I will provide two backup meeting systems. In case there are technical problems with Zoom, you should switch and join any one of the two alternatives immediately.
  + Google meet (To join the video meeting, click this link: <https://meet.google.com/test>, you don’t need to install anything)
  + Tencent Meeting (腾讯会议会议: 72746test, you need to install the app in advance).
* Final exam duration is 3 hours. 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.
* Your solutions should be compiled in one word file (copy all your R codes into the box provided), and be submitted to Canvas by the due time. *Late submission:* ***10% of your scores will be deducted***. ***The exam submission link will be disabled at 21:45pm. Submission after that will not be accepted.***
* No collaboration or online communications is allowed.
* In case of emergency, contact me at +852 3442-8650; The departmental hotline is 3442-8325.

**Final Practice Exam**

**Question 1. True/False**

Mark each statement True or False. If you believe that a statement is false, briefly explain why you think it is false.

(a) The presence of collinearity violates an assumption of the Multiple Regression Model (MRM).

(b) The F-statistic is statistically significant, i.e., the whole model is significant, only if some t-statistic for a slope in multiple regression is statistically significant.

**Question 2.**

**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).

(a) 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)

(b) Fit the indicated multiple regression and show a summary of the estimated model.

(Attach the regression outputs here)

(c) Does this estimated model explain statistically significant variation in the cost of the leases?

(d) 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.

(e) 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 here)

(f) Interpret the estimated coefficient for the interaction between *Square.Feet* and *Location* of the model in part (e). What does the estimated value mean?

**Question 3.**

**admission.csv**

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, affect admission into graduate school. The response variable*,* ***admit***, is a binary variable (admit = 0 means the student got admitted by the graduate school, admit = 0 means the student got rejected). The explanatory variables include

* ***gre***: the GRE score (numerical variable)
* ***gpa***: grade point average (numerical variable)
* ***rank***: the rank of the prestige of the undergraduate institution (categorical variable with four levels)

(a) Run a logistic regression to predict the probability of admission to graduate school using all the given predictors excluding their interactions. **Make sure that *rank is a* categorical variable.**

(b) With the model in part (a), predict the probability of admission for the student with gre=600, gpa=3.1, and rank=2.

(c) The logistic regression provides a probability of being admitted given the explanatory variables of each student. Given a threshold, if the predicted probability is greater than it, we say the predicted admission is 1, or ***admit***, otherwise, 0 or ***reject***.

We next define two concepts:

Sensitivity is also called the **true positive rate** or **probability of detection.** Itmeasures the proportion of actual positives that are correctly identified as such (e.g., the percentage of admit students 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 reject students who are correctly identified as being rejected).

Let the threshold be 0.5. Use R to compute the sensitivity and specificity. Possible useful R functions:

cbind(), table(), nrow()