P8130: Biostatistical Methods I
Final Project (Fall 2017)
Due, December 15th @ 12:00pm

Guidelines for Project Submission

This group project must be submitted through Canvas before the deadline. Late email submissions WILL NOT be accepted and will receive a score of 'Zero' for all group members!!

All graphs, output and interpretations must be included in **ONE PDF** (not the R/SAS code), otherwise it will not be graded. You also have to submit the R/SAS code used in your project, but please add it at the end of the document or in a separate attachment.

General Writing Instructions

The PDF containing your summary should not exceed **5 double-spaced pages** using 11 or 12-point font, INCLUDING figures and tables, references, appendix, etc. However, you do not need to fill all 5 pages – a succinct, but comprehensive summary will be very much appreciated.
Your report should be structured as a publishable research article containing the following sections:

- Abstract,
- Introduction (context, background of the problem),
- Methods (data description and statistical methods),
- Results,
- Conclusions/Discussion

In the model, write the full name rather than variable name

P–values, confidence intervals

Journal like statistics in medicine, JCO

Your findings should be written as for an informed (but non-statistical) audience (no formulae!). Each figure and table should be of publishable quality and well notated, i.e., labeled and/or captioned.

Grading Instructions

The rubric attached will be used to evaluate the project. This is a group project and collaborations within your group are essential. Therefore, 90% of your individual grade will be based on the group report, and 10% on group participation. Each member of the group will review his/her colleagues and give them a score from 1 (worst) to 5 (best). I will then average the ratings and use that score for your 10% group participation.

Academic dishonesty will be punished with a 'Zero' grade for this project.

Linear regressions to pick up variables, criteria–based to determine final model. Find other study, are there any other significant variables. Do frequency table, histagram. Keep every categories? Collapse categories, combine them if clinically make sense.

The Data Analytics group from Good Health Corporation are interested in improving the overall hospital management and minimizing the cost/resources associated with patients' care. One of the most important outcomes that has a direct effect on these aspects is patient's length of stay (LoS) in the hospital. Thus, they would like to know which variables are associated with LoS, and ultimately build a predictive model to be used for future visits. The group has contacted you to study this problem and make a recommendation.

Data Description:

A total of 3682 records from 3612 patients were collected in the 2016 calendar year. Only visits within 24 hours of hospital admission and for patients older than 17 years were considered relevant for this analysis. Below you have a selected list of variables that **need to be considered in your analysis.** Feel free to explore other variables as well, but make sure that you first address the ones below.

PatientID:	unique identifier for each patient
VisitID:	unique number for each admission (can be multiple per PatientID)
AdmitDtm:	complete date of admission
LOSDays2:	length of stay in the hospital (days)
Ls30DayReadmit:	1=admission into the hospital within past 30 days; 0=otherwise.
MEWS:	The Modified Early Warning Score (MEWS) determines the degree of illness of a patient based on respiratory rate, oxygen saturation, temperature, blood pressure, heart rate, AVPU response; 0-1=normal, 2-3=increase caution, 4-5=further deterioration, >5 immediate action required
Cindex:	Charlson comorbidity index (CCI) ranks patients based on severity of comorbidity: 0=normal, 1-2=mild, 3-4=moderate and >5=severe
Evisit:	number of times the patient visited an emergency department in the six months prior to admission (not including the emergency department visit immediately preceding the current admission)
ICU_Flag:	1=if during hospitalization, the patient had a visit in the intensive care unit (ICU); 0=otherwise. Note that ICU patients tend to have more hospital related conditions and thus a longer length of stay.
AgeYear:	patient's age in years
Gender:	patient's gender
Race:	patient's race
Religion:	patient's religion
MaritalStatus:	patient's marital status
InsuranceType:	patient's insurance

Frequency table, might collapse 4 to 2 categories if not significant. Is it normal

Can we discard several ICU, since the one with ICU visits tends to prolong the LOS

Vital Signs: respiration rate, blood pressure diastolic (BPD), oxygen saturation (O2), blood pressure diastolic (BPS), temperature, heart rate, and body mass index (BMI).

Note: Each VisitID represents a unique visit. However, it is possible that a patient visited the hospital more than once. Summarize the number of visits per patient and if multiple visits per patient, select the first visit (by date) recorded.

(Some) things you should consider:

- Data cleaning: replace any 'funny' characters with missing values
- Transformation(s) of the outcome
- Re-code/combine levels of categorical variables based on frequency and practical importance
- Check the predictive capability of the model Cross validation or Bootstrap
- In this course, we only covered linear regression models. Let us assume that even after exploring different combinations of predictors your model does not fit the data well and/or does not have a good predictive ability. What other statistical methods/models not covered in this course would you recommend for future steps?