NATIONAL UNIVERSITY OF SINGAPORE

ST3243 STATISTICAL METHODS IN EPIDEMIOLOGY

Project Assignment, AY2016/2017

INSTRUCTIONS

1. This project consists of TWO independent sections.

2. The Section I is a practical section which each student is given for a unique dataset (can be found in IVLE website). Each dataset consists of 400 records that randomly re-sampled from the original dataset (N=200). The data description and the questions to this dataset are attached in the Section 1A.

3. Each student should independently perform data analysis and apply appropriate statistical methods to answer the questions.

4. The Section II would contain 3 questions that required written answer only, but please note that answer to each of these questions should not be longer than 2 pages.

5. The project report (in softcopy will do) should be completed, submitted and uploaded to IVLE website by **11.59pm of Sunday 06 November 2016**.

6. Your assignment report should be named under with both of your Full Name and Matriculation Number (e.g. Tan Tai Beng_a123456E).

7. The project report should include clear answers one by one to each question and attach with statistical programming codes (where appropriate) using R or any other statistical package that you are most familiar with.

8. This project assignment would account for 35% of the total marks for this ST3243 course.

Section I

A. Data description

The intensive care unit (ICU) Study

The ICU data set consists of a re-sample of 400 subjects who were part of a much larger study on survival of patients following admission to an ICU. The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients.

A code sheet for the variables to be considered in this test is given in the following table.

Variable	Description	Codes/Values	Name
1	Identification Code	ID Number	ID
2	Vital Status	0 = Lived,1= Died	STA
3	Age	Years	AGE
4	Sex	0 = Male1 = Female	SEX
5	Race	1 = White, 2 = Black, 3 = Other	RACE
6	Service at ICU Admission	0 = Medical, 1 = Surgical	SER
7	Cancer Part of Present Problem	0 = No, 1 = Yes	CAN
8	History of Chronic Renal Failure	0 = No, 1 = Yes	CRN
9	Infection Probable at ICU Admission	0 = No, 1 = Yes	INF
10	CPR Prior to ICU Admission	0 = No, 1 = Yes	CPR
11	Systolic Blood Pressure at ICU Admission	mm Hg	SYS
12	Heart Rate at ICU Admission	Beats/min	HRA
13	Previous Admission to an ICU Within 6 Months	0 = No, 1 = Yes	PRE
14	Type of Admission	0 = Elective, 1 = Emergency	TYP
15	Long Bone, Multiple, Neck, Single Area, or Hip Fracture	0 = No, 1 = Yes	FRA
16	PO2 from Initial Blood Gases	0 = >60, 1 = <=60	PO2
17	PH from Initial Blood Gases	0 = >=7.25, 1 = <7.25	PH
18	PCO2 from Initial Blood Gases	0 = <=45, 1 = >45	PCO
19	Bicarbonate from Initial Blood Gases	0 = >=18, 1 = <18	BIC
20	Creatinine from Initial Blood Gases	0 = <=2, 1 = >2	CRE
21	Level of Consciousness at ICU Admission	0 = No Coma or Deep Stupor, 1 = Deep Stupor, 2= Coma	LOC

B. Questions

1. In the ICU data described above, the primary outcome variable is vital status at hospital discharge, STA to be used throughout in this assignment. Clinicians associated with the study felt that a key determinant of survival was the patient's age at admission, AGE.

    a. Write down the equation for the logistic regression model of STA on AGE. Write down the equation for the logit transformation of this logistic regression model. What characteristic of the outcome variable, STA, leads us to consider the logistic regression model as opposed to the usual linear regression model to describe the relationship between STA and AGE?
    b. Form a scatter-plot of STA versus AGE.
    c. Using the intervals [15, 24], [25, 34], [35, 44], [45, 54], [55, 64], [65, 74], [75, 84], [85, 94] for AGE, compute the STA mean over subjects within each AGE interval. Plot these values of mean STA versus the mid-point of the AGE interval using the same set of axes as was used in Question 1(b).
    d. Using R or any other statistical package to obtain estimate of the parameters of the logistic regression model in Question 1(a). These estimates should be based on the ungrouped, n=400, data. Using these estimates, write down the equation for the fitted values, that is the estimated logistic probabilities. Plot the equation for the fitted values on the axes used in the scatter-plots in Questions 1(b) and 1(c).
    e. Summarize (describe in words) the results presented in the plot obtained from Questions 1(b), 1(c), and 1(d).
    f. Using the results of the output from the R or any other statistical package used for Question 1(d), assess the significance of the slop coefficient for AGE.
    g. Using the results from Question 1(d) compute 95% Confidence Intervals (CI) for the slop and intercept term. Write a sentence interpreting the CI for the slope.
    h. Compute the logit and estimated logistic probability for a 60-year old subject. Compute a 95% CI for the logit and estimated logistic probability. Write a sentence or two interpreting the estimated probability and its CI.
    i. Using the R or any other statistical package to obtain the estimated logit and its standard error for each subject in the ICU study. Graph the estimated logit and the point-wise 95% confidence limits versus AGE for each subject. Explain (in words) the similarities and differences between the appearance of this graph and a graph of a fitted linear regression model and its point-wise 95% confidence bands.

2. Using the outcome variable vital status (STA) and CPR prior to ICU admission (CPR) as a covariate.

    a. Demonstrate that the value of the log-odds ratio obtained from the cross-tabulation of STA by CPR is identical to the estimated slope coefficient from the logistic regression of STA on CPR. Verify that the estimated standard error of the estimated slope coefficient for CPR obtained from the R or other statistical package is identical to the square root of the sum of the inverse of the cells frequencies from the cross-tabulation of STA by CPR. Use either set of computations to obtain 95% CI for the odds ratio. What aspect concerning the coding of the variable CPR makes the calculations for the two methods equivalent?
    b. For purpose of illustration, use a data transformation statement to recode, for this problem only, the variable CPR as follows: 4=no and 2=yes. Perform the logistic regression of STA on CPR (recoded). Demonstrate how the calculation of the logit difference of CPR=yes versus CPR=no is equivalent to the value of the log-odds ratio obtained in the Question 2(a). Use the results from the logistic regression to obtain the 95% CI for the odds ratio and verify that they are the same limits as obtained in Question 2(a).
    c. Consider the ICU data and use as the outcome variable vital status (STA) and race (RACE) as covariate. Prepare a table showing the coding of the two dummy variables for RACE using the value RACE=1, white as the reference group. Show that the estimated log-odds ratios obtained from the cross-tabulation of STA by RACE, using RACE=1 as the reference group, are identical to estimated slope coefficients for the two dummy variables from the logistics regression of STA on RACE. Verify that the estimated standard errors of the estimated slope coefficients for the two dummy variables for RACE are identical to the square root of the sum of the inverse of the cell frequencies from the cross-tabulation of STA by RACE used to calculate the odds ratios. Use either set of computations to compute the 95% CI for the odd ratios.
    d. Prepare a table showing the coding of three dummy variables based on the empirical quartiles of AGE using the first quartiles as the reference group. Fit the logistic regression of STA on AGE as recoded into these design variables and plot the three estimated slope coefficients versus the mid-point of the respective age quartile. Plot as a fourth point a value of zero at the mid-point of the first quartile of age. Dose this plot suggest that the logit is linear in age?
    e. Consider the logistics regression of STA on CRN and AGE. Consider CRN to be the risk factor and show that AGE is a confounder of the association of CRN with STA. Addition of the interaction of AGE by CRN presents an interesting modelling dilemma. Examine the main effects only and interaction models

graphically. Using the graphical results and any significance tests you feel are needed, select the best model (main effects or interaction) and justify your choice. Estimate relevant odds ratios. Repeat this analysis of confounding and interaction for a model that includes CPR as the risk factor and AGE as the potential confounding variable.

f. Consider an analysis for confounding and interaction for the model with STA as the outcome, CAN as the risk factor, and TYP as the potential confounding variable. Perform this analysis using logistic regression modelling.

Section II

1. Describe what are the differences between odds ratio and relative risk, in terms of study design, interpretations etc.

2. Describe three methods that commonly used in epidemiological data analysis to control confounding effects.

3. Describe the two criteria to assess the adequacy of a logistic regression model.