Homework 2

BSTA 513/613

Your name here - update this!!!!

> **Caution**
>
> This homework is ready to go! (Nicky 4/17/25)

Purpose

This homework is designed to help you practice the following important skills and knowledge that we covered in Lessons 6-9:

- Test a covariate for significance using the Wald test and LRT
- Fit and interpret simple and multivariable logistic regression models
- Interpret odds ratios and their confidence intervals
- Evaluate model fit using likelihood-based methods
- Apply data wrangling techniques to prepare variables for regression modeling
- Communicate results of logistic regression analyses clearly and accurately in writing

Directions

- Download the `.qmd` file here.
- You will need to download the datasets from our shared folder.
- Please upload your homework to Sakai. Upload both your .qmd code file and the rendered .html file
  - Please rename you homework as Lastname_Firstinitial_HW02.qmd. This will help organize the homeworks when the TAs grade them.
- For each question, make sure to include all code and resulting output in the html file to support your answers

- Show the work of your calculations using R code within a code chunk. Make sure that both your code and output are visible in the rendered html file. This is the default setting.
- Write all answers in complete sentences as if communicating the results to a collaborator.

> Tip
>
> It is a good idea to try rendering your document from time to time as you go along! Note that rendering automatically saves your qmd file and rendering frequently helps you catch your errors more quickly.

Questions Part 1

The following questions are intended to give you **practice in understanding concepts** and **completing calculations**.

Question 1

This question is taken from the Hosmer and Lemeshow textbook. The ICU study data set consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The dataset should be available in our shared folder. 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. In this question, the primary outcome variable is vital (survival) status at hospital discharge, STA. Clinicians associated with the study felt that a key determinant of survival was the patient's age at admission, AGE.

A code sheet for the variables to be considered is displayed in Table 1.5 below (from the Hosmer and Lemeshow textbook, pg. 23). We refer to this data set as the ICU data.

Table 1.5 Code Sheet for the Variables in the ICU Data

Variable	Description	Codes/Values	Name
1	Identification code	ID number	ID
2	Vital status at hospital discharge	0 = Lived 1 = Died	STA
3	Age	Years	AGE
4	Gender	0 = Male 1 = Female	GENDER
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	TYPE
15	Long bone, multiple, neck, single area, or hip fracture	0 = No 1 = Yes	FRA
16	PO_2 from initial blood gases	0 = >60 1 = ≤60	PO2
17	PH from initial blood gases	0 = ≥7.25 1 = <7.25	PH
18	PCO_2 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 = ≤2.0 1 = >2.0	CRE
21	Level of consciousness at ICU admission	0 = No coma or deep stupor 1 = Deep stupor 2 = Coma	LOC

Part a

Write down the equation for the population logistic regression model of STA on AGE. 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?

Part b

Write down an expression for the log-likelihood for the logistic regression model in Part a. This will be a mathematical expression. Please do not use generic expressions like $\pi(X)$, instead replace X with the specific variables in this question.

Part c

Using the `glm()` function, obtain the maximum likelihood estimates of the coefficient parameters of the logistic regression model in Part a. Using these estimates, write down the equation for the fitted logistic regression model.

Part d

Use the Wald test to test whether or not the intercept (β_0) of the logistic regression model is significantly different from 0. Make sure to follow the steps laid out in class and include: hypothesis test, code/work leading to the computed test statistic, output including the test statistic and p-value, and conclusion.

Part e

Use the Likelihood Ratio test to test whether or not the coefficient for age (β_1) of the logistic regression model is significantly different from 0. Make sure to follow the steps laid out in class and include: hypothesis test, code/work leading to the computed test statistic, output including the test statistic and p-value, and conclusion.

Part f

Write a sentence interpreting the estimated odds ratio for the coefficient in Part e. Please include the 95% confidence interval.

Question 2

We will continue to work with the ICU data in Question 1. Please refer back to the information above. In this question, we will use the ICU data to fit a multivariable logistic regression model.

Part a

From the above list (AGE, CAN, CPR, INF, and LOC) of independent variables, identify if each is a continuous, binary, or multi-level (>2) categorical variable.

Part b

For the binary and multi-level categorical variables, please identify a reference group for each. Include justification for the reference group.

Part c

Compute the predicted probability of hospital discharge for a subject who is 63 years old. Compute the 95% confidence interval for the predicted probability and interpret the predicted probability.

Part d

For the categorical variables (binary and multi-group), please mutate the variables within the ICU dataset to set your chosen reference groups.

Part e

Write down the equation for the logistic regression model of STA on CPR.

Part f

Using the `glm()` function, obtain the maximum likelihood estimates of the coefficient parameters of the logistic regression model in Part f. Using these estimates, write down fitted logistic regression model.

Part g

Write a sentence interpreting the odds ratio for the coefficients in Part g's model. Please include the 95% confidence interval.

Part h

Write down the equation for the logistic regression model of STA on LOC.

Part i

Using the `glm()` function, obtain the maximum likelihood estimates of the coefficient parameters of the logistic regression model in Part h. Present the coefficient estimates. No need to write out the fitted regression equation.

Please take note of the warnings that you receive from fitting the `glm()` model and any large coefficient estimate with large confidence intervals. In this case, we have a category within LOC that has very few observations. (We will discuss this more in Lesson 14: Numerical Problems)

Check the number of observations that have a deep stupor and death at discharge and the number of observations that have a deep stupor and live at discharge. You can do this using the `table()` function to create a contingency table.

Part j

Write a sentence interpreting the odds ratio of death for the indicator of coma. Please include the 95% confidence interval.

Questions Part 2

The following questions are intended to give you **practice in connecting concepts** that will help you make decisions in real world applications.

Question 3

Using a similar table to the one in Lesson 8, go back through the parts in this homework and determine which test can be run .

	Wald test	Score test	LRT
Question 1, Part d: testing intercept			
Question 1, Part e: testing coefficient(s) for age			

	Wald test	Score test	LRT
Question 2, Part e/f: testing coefficient(s)for CPR			
Question 2, Part h/i: testing coefficient(s)for LOC			