Categorical Data Analysis, Prelim 2

STSCI/BTRY 4110 ---- November 7, 2023

Cover Sheet

Name _____

Net ID _____

Academic integrity is expected of all students of Cornell University at all times, whether in the presence or absence of members of the faculty.

Understanding this, I declare I shall not give, use, or receive unauthorized aid in this examination.

Signatures of Students _____

Notes:
> This cover sheet should be the first page of your completed report.

> You should submit an electronic copy of your work via Canvas. The electronic copy may be checked for plagiarism.

> Your submission should be two files: 1) Your written report, 2) R code for the analysis. You do NOT need to include all the R code that you ran, just the R code that generated the analyses described in your report.

> In addition, you need to fill out and submit a Peer Group Evaluation Form. This is required for everyone. This is due 24 hours after the project has been submitted. Your comments on teammates will be kept confidential. Only Dr. Smith will see the completed forms.

> The report should be the work of you and your partners, written in your own words. Do not discuss the exam with other classmates or students, professors or other people. You may ask me questions. It should be no longer than 12 pages long.

> The deadline for submission of your report and R code is Friday November 17 at 11:59 pm.

Analysis of Cardiac Events

Effects of Sleep and Body Mass Index (BMI)

In a national cross-sectional study of health status in the United States, a random sample of adults aged 20 and above were asked questions about their current health status.

The data in this study were collected in the time period stretching from 2017 to March 2020 from 1910 individuals. Demographic data on age, gender, race and education were collected, in addition to measures of health: sleep hours per night, diabetes status, smoking status and bmi. The outcome variable of interest was the presence or absence of a cardiac event. A person was classified as having a cardiac event if a doctor had told them that they had or have congestive heart failure, angina, coronary heart disease, hypertension, MI ("heart attack") or stroke. The data file is named "final_cardiac_data.csv".

The objective of this assignment is two-fold:

1) You should analyze and determine which factors might be predictive of cardiac events. You should also evaluate and quantify how those factors increase or decrease the risk of a cardiac event.

2) As part of this analysis, you should specifically address whether or not there is evidence of an effect of nightly sleep hours on cardiac events. In this analysis, you should determine the effect of sleep hours, when adjusting for other factors. Note that you should consider whether sleep hours should be used as a numerical continuous variable or as a categorical variable with 3 levels (<= 6, 6-8.9, >=9) or as a categorical variable with 2 levels (<=6, >6)

The Codebook has an explanation of each variable and its coding in the dataset. It is given at the end of this document. Note that your should recode the race variable in such a way that there are just three categories: "White", "Black" and "Other".

Your assignment is to write a report on the factors influencing cardiac events. Health and sleep researchers and members of the public want to know which factors or combination of factors are most predictive of cardiac events. They are particularly interested in whether or not sleep has an effect on the incidence of cardiac events, and, if so, how strong is the effect of sleep. If the effect of sleep depends on the levels of the other variables, you should indicate this too.

Your report should include the fitting of a logistic regression model and a description of the steps you took to fit the model. (Note that you should also include other types of categorical data analysis tools such as contingency analysis for preliminary analyses before fitting your model.) Your write-up should include graphical displays, as appropriate.

Use the *Take Home Project Guidelines* posted on Canvas as a guide to my expectations. The Guidelines provide a detailed list of all items that should be included in your written report.

If there are situations during the analysis when you would want to discuss a next step with a doctor or health researcher for advice, mention this in your report. Make any such decisions, as best you can.

You do not have to include a description of every single thing you did or every step that you took; you should outline the steps that you took, with relevant intermediate results. Your report should have more detail than you would find in a typical published paper, but less than a step-by-step description.

If you have questions, you may post a private message to me in Ed or email me directly at ms429@cornell.edu. If the questions are relevant to everyone, I will post them on Ed Discussions with answers. Likewise, if I have clarifications or corrections, I will post them on the Ed discussion board.

Good luck and enjoy!
Melissa Smith
(ms429)

Codebook

Cardiac event "event"

Has a doctor ever told you that you have or had: congestive heart failure, angina, coronary heart disease, hypertension, MI ("heart attack") or stroke?

1 = Yes, 0 = No

Gender "gender"

Code or Value	Value Description
1	Male
2	Female

Education "educ"

What is the highest grade or level of school that you have completed?

Code or Value	Value Description
1	Less than 9th grade
2	9-11th grade (Includes 12th grade with no diploma)
3	High school graduate/GED or equivalent
4	Some college or AA degree
5	College graduate or above
7	Refused
9	Don't Know

Race/Hispanic Origin "ethnic1"

You should recode this to three categories: White, Black, Other

Code or Value	Value Description
1	Mexican American
2	Other Hispanic
3	Non-Hispanic White
4	Non-Hispanic Black
5	Other Race - Including Multi-Racial

Diabetes “diabetes”

Code or Value	Value Description
1	Yes
2	No
3	Borderline
7	Refused
9	Don't know

Smoker “smoker”

Have you used any tobacco product in the last 5 days?

Code or Value	Value Description
1	Yes
2	No
7	Refused
9	Don't know

Age “age” (in years)

BMI “bmi” is Body Mass Index measured in kg/m^2

BMI can be categorized into 4 categories:

"Underweight": <=18.5

“Normal”: 18.6 - 24.9

“Overweight”: 25-29.9

"Obese": >= 30 See

See the section on helpful R code

Sleep “sleep.hrs” How many hours of sleep do you get on weekdays or workdays?

Some Examples of Recoding in R

```
card <- read.csv("final_cardiac_data.csv")
card$age.cat <- cut(card$age,
                    breaks=c(-Inf, 40, 55, 70, Inf),
                    labels=c("21-40","41-55","56-70", "70+"))
table(card$age.cat, card$age)

card$bmi.cat <- cut(card$bmi,
                   breaks=c(-Inf, 18.5, 24.9, 29.9, Inf),
                   labels=c("Underweight","Normal","Overweight", "Obese"))

card$gcode[card$gender==1] <- "Male"
card$gcode[card$gender==2] <- "Female"
```