BIOST 536 Homework 4

Benjamin Stan

October 30, 2021

Question 1

The results of a logistic regression show that the odds ratio of esophageal cancer between two populations of the same age group defined using grouped-linear adjustment, one that consumes more than 10g/day of cider and one that does not, is 2.19 (95% robust CI: 1.55-3.08), with the cider-consuming group having increased odds.

Question 2

The results of a logistic regression show that the odds ratio of esophageal cancer between two populations of the same age group defined using indicator variables, one that consumes more than 10g/day of cider and one that does not, is 2.03 (95% robust CI: 1.44-2.85), with the cider-consuming group having increased odds.

Question 3

a.) The results from Q1 and Q2 differ surprisingly given the similarity in their interpretations. The grouped linear approach generated an estimated odds ratio of 2.19 (95% robust CI: 1.55-3.08) adjusted for age, while the indicator variable approach generated an estimated odds ratio of 2.03 (95% robust CI: 1.44-2.85). While there is substantial overlap in their confidence intervals, there is a difference in point estimate and, in particular, the upper bound of their confidence intervals.

b.) Between the two methods used in Q1 and Q2, I would prefer the indicator variable approach. This is because this method is the most thorough one which uses grouped versions of a continuous variable and does not impose a uniform value corresponding to the increase in age from one group to the next highest.

Question 4

Using the results of a logistic regression, the odds ratio of esophageal cancer between two populations of the same age, one that consumes more than 10g/day of cider and one that does not, is 2.17 (95% robust CI: 1.54-3.05), with the cider-consuming group having increased odds.

Question 5

Using the results of a logistic regression, the odds ratio of esophageal cancer between two populations of the same first and second orders of age (age and age^2), one that consumes more than 10g/day of cider and one that does not, is 1.97 (95% robust CI: 1.40-2.77), with the cider-consuming group having increased odds.

Question 6

A table show summary information for the data set is below:

agegp	n	agemin	agemax	prop_cancer
1	116	25	34	0.009
2	199	35	44	0.045
3	213	45	54	0.216
4	242	55	64	0.314
5	161	65	74	0.342
6	44	75	91	0.295

a.) The table above shows the age groups defined in the data set and used for the linear spline regression. As can be seen, the proportion of subjects with cancer increases with increasing age group, and there is no single age group that is low in sample size (which ranges from 44-242 per age group). For this reason, knots beginning at 35 and increasing by 10 until 75 will be used.

b.) The model under consideration is a logistic regression on the probability of esophageal cancer with age adjustment using linear splines and knots at ages 35, 45, 55, 65, and 75. The results show that the odds ratio of esophageal cancer between two populations of the same age, one that consumes more than 10g/day of cider and one that does not, is 1.98 (95% robust CI: 1.41-2.78), with the cider-consuming group having increased odds.

Question 7

a.) The odds ratio estimates along with their 95% robust confidence interval bounds are shown in the table below:

Method	Point Estimate (e^{β_1})	CI Lower	CI Upper
Continuous Linear (Q4)	2.17	1.54	3.05
Continuous Quadratic (Q5)	1.97	1.40	2.77
Linear Spline (Q6)	1.98	1.41	2.78

In comparing these values, it appears that the continuous quadratic and linear spline approaches yield very similar results. However, these two approaches differ from the results of the continuous linear adjustment. This is consistent with the intuition that the continuous quadratic and linear spline approaches give flexible fits that can conform more to the patterns of the data than a simple linear fit.

b.) The approach that I would choose if performing this analysis is the linear spline approach, as it provides a rigorous yet flexible solution to address the confounding variable. It also is provides an interpretable solution, as it could be sensible that the linear relationship between disease and age differs depending on the age range under consideration; this is an area where a single quadratic fit suffers.

Question 8

A summary table of the models is shown below:

Question	Approach	Model
1	Grouped Linear	$logit(p) = \beta_0 + \beta_1 * cider + \alpha * agegp$
2	Indicator Variables	$logit(p) = \beta_0 + \beta_1 * cider + \sum_{i=1}^{5} \alpha_i * I_{agegp=i}$
4	Continuous Linear	$logit(p) = \beta_0 + \beta_1 * cider + \alpha * age$
5	Continuous Quadratic	$logit(p) = \beta_0 + \beta_1 * cider + \alpha_1 * age + \alpha_2 * age^2$
6	Linear Spline	$logit(p) = \beta_0 + \beta_1 * cider + \alpha_1 * age + \sum_{i=2}^{6} \alpha_i * s_i$

a.) The Q1 model is nested in Q2 because the model in Q1 could be generated by setting the difference between all successive indicator coefficients in the Q2 model to be the same (i.e. $\alpha_1 = \alpha_2 - \alpha_1 = \alpha_3 - \alpha_2 = \alpha_4 - \alpha_3 = \alpha_5 - \alpha_4$). The Q2 model is not nested in Q1 because there is no way to remove the uniform difference associated with an agegp increase of one from the Q1 model.

b.) The models in Q1 and Q4 are not nested in either direction because the difference in age for an agegp difference of one is not consistent.

c.) The model in Q4 is nested within Q5, which can be seen by setting the quadratic term in the Q5 model to zero. The Q5 model is not nested within the Q4 due to the quadratic term.

d.) The model in Q4 is nested within the Q6 model, which can be seen by setting all of the coefficients on the spline variables equal to zero (i.e. $\alpha_2 = \ldots = \alpha_6 = 0$). The presence of these spline variables, however, makes it such that the Q6 model is not nested within Q4 model.

e.) The models in Q5 and Q6 are not nested in either direction, as they each have terms that the other does not have. These are the quadratic term in Q5 and the spline terms in Q6.

Appendix

```
## Question 1
setwd("/Users/bstan/Documents/UW/Courses/BIOST 536")
rm(list = ls())
library(tidyverse)
library(tidyr)
library(tinytex)
library(sandwich)
load("data/esophcts.Rdata")

### Fit logistic regression
esophcts <- esophcts %>% mutate(cider_bin = ifelse(cider>10,1,0))
glm <- glm(case ~ cider_bin + agegp, data = esophcts, family = "binomial")
glm$coef %>% exp

coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int <- coef[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se[2]
conf_int %>% exp

## Question 2
```

```r
glm <- glm(case ~ cider_bin + as.factor(agegp), data = esophcts, family = "binomial")
glm$coef %>% exp

coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int <- coef[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se[2]
conf_int %>% exp

## Question 4
glm <- glm(case ~ cider_bin + age, data = esophcts, family = "binomial")
glm$coef %>% exp

coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int <- coef[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se[2]
conf_int %>% exp

## Question 5
glm <- glm(case ~ cider_bin + age + I(age^2), data = esophcts, family = "binomial")
glm$coef %>% exp

coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int <- coef[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se[2]
conf_int %>% exp

## Question 6
### Explore the range of each age group
knitr::kable(esophcts %>% group_by(agegp) %>%
  summarise(n=n(),
            agemin = min(age),
            agemax=max(age),
            prop_cancer=round(sum(case)/n(),3)))

### Fit the logistic regression
glm <- glm(case ~ cider_bin + age_spline + s1 + s2 + s3 + s4 + s5,
           data = esophcts, family = "binomial")
glm$coef %>% exp

coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int <- coef[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se[2]
conf_int %>% exp
```