Homework 4

David Coomes

11/4/2019

1. The OR of being diagnosed with esophageal cancer associated with consuming more than 10 grams of cider per day, after controlling for age using grouped-linear adjustment, is 2.19 (95% CI: 1.56-3.07).

	or	2.5 %	97.5 %
bin_cider	2.19	1.56	3.07

2. After controlling for age using categorical dummy variables spanning 10 years, the OR of being diagnosed with esophageal cancer associated with more than 10 g/day of cider as compared to less than 10 g/day is 2.03 (95% CI: 1.44-2.85)

	or	2.5 %	97.5 %
bin_cider	2.03	1.44	2.85

3. (a) The findings from question 1 and 2 are materially different because, although the results are in the same direction, when adjusting for age using grouped-linear adjustment the OR is about 20% higher as compared to adjusting using dummy variables when comparing them to the null. I believe this is a meaningful difference.

   (b) I prefer the dummy variable adjustment because it does not constrain the model as much (it allows it to better fit the true data), and becuase our data are rich enough to support it - especially without other confounding variables.

4. The chi-square statistic for the LRT between the reduced (case and age) and full (case, cider consumption, and age) models are 86.594 and the p-value of this test statistic is less than 2.2 * 10^-16

	or	2.5 %	97.5 %
bin_cider	2.17	1.54	3.04

Using case-control data, we fit a logistic regression model with diagnosis with esophageal cancer as the outcome variable, a binary variable of consumption of more than 10 g/day of cider as the exposure variable, and controlling for age as a continuous variable.

5. The chi-square value for the LRT between the reduced (case and age, and age-squared) vs. the full (case, cider, age, and age-squared) model is 15.525 and the p-value of this statistic is 8.142 * 10^-5

	or	2.5 %	97.5 %
bin_cider	1.97	1.4	2.77

Using case control data, we fit a logistic regression model using diagnosis with esophageal cancer (case) as the outcome variable, and cider consumption as a binary variable (more than 10 g/day) as the exposure variable, while controlling for age quadratically by including age and age-squared in the model.

6. The chi-square value for the LRT between the reduced (case and age) vs. the full (case, cider, age) model is -15.788 and the p-value of this statistic is 7.085 * 10^-5

	or	2.5 %	97.5 %
bin_cider	1.98	1.41	2.79

Using case-control data, we fit a logistic regression model using diagnosis with esophageal cancer as the outcome of interest, cider consumption (binary variable >10 g/day) as the main exposure variable, and age as a linear spline adjustment variable using 10-year age ranges beginning at age 25.

7. The OR estimates and 95% CIs are very similar between the quadratic and linear spline age adjustments. These ORs are slightly lower and the 95% CIs are slightly smaller as compared to the linear age adjustment model. I think that either the quadratic or the linear spline adjustments are the preferable methods becuase they fit the data better and allow for a richer model. Since the estimates are very similar and the data is rich enough, I think that either of those models are okay, although I may choose the quadratic because it uses less adjustment variables and is likely more accurate because it allows for smoother variation within age groups.

8. Comparing the dummy model to the quadratic model, the results, including the OR and the 95% CI are very similar. I would argue that there is no meaningful difference between the two. If I had to choose one, I may go with the quadratic model because it seems to fit the data well and allows for variation within the age categories. Although I think that either one would be fine in a study.

Appendix

```r
library(knitr)
library(devtools)
library(UWbe536)
library(kableExtra)
library(formatR)


link = "https://github.com/dmccoomes/Biostats_536/raw/master/Homework%204/esophcts.rds"
esoph <- readRDS(url(link))
head(esoph)
summary(esoph)
str(esoph)


esoph$bin_cider[esoph$cider <= 10] <- 0
esoph$bin_cider[esoph$cider > 10] <- 1


esoph$age1 <- as.numeric(esoph$agegp)
cider.mod <- glm(case ~ bin_cider + age1, data = esoph, family = binomial)
# coef(summary(cider.mod))
or <- exp(coef(cider.mod)[2])
ci <- exp(confint.default(cider.mod)[2, ])

kable(round(cbind(or, t(ci)), 2)) %>% kable_styling(full_width = F,
    position = "center")




esoph$age2 <- as.factor(esoph$agegp)
cider.mod_dum <- glm(case ~ bin_cider + age2, data = esoph, family = binomial)
# coef(summary(cider.mod_dum))

or <- exp(coef(cider.mod_dum)[2])
ci <- exp(confint.default(cider.mod_dum)[2, ])
```

```
kable(round(cbind(or, t(ci)), 2)) %>% kable_styling(full_width = F,
    position = "center")
```

```
cider.mod_con <- glm(case ~ bin_cider + age, data = esoph, family = binomial)
# coef(summary(cider.mod_con))

cider.mod_red <- glm(case ~ bin_cider, data = esoph, family = binomial)

# running LRT for reduced vs. full model anova(cider.mod_red,
# cider.mod_con, test='LRT')

or <- exp(coef(cider.mod_con)[2])
ci <- exp(confint.default(cider.mod_con)[2, ])

kable(round(cbind(or, t(ci)), 2)) %>% kable_styling(full_width = F,
    position = "center")
```

```
cider.mod_con <- glm(case ~ bin_cider + age, data = esoph, family = binomial)
# coef(summary(cider.mod_con))

esoph$age_2 <- esoph$age^2
cider.mod_quad <- glm(case ~ bin_cider + age + age_2, data = esoph,
    family = binomial)

cider.mod_quad_red <- glm(case ~ age + age_2, data = esoph, family = binomial)

# running LRT for reduced vs. full model
# anova(cider.mod_quad_red, cider.mod_quad, test='LRT')

or <- exp(coef(cider.mod_quad)[2])
ci <- exp(confint.default(cider.mod_quad)[2, ])

kable(round(cbind(or, t(ci)), 2)) %>% kable_styling(full_width = F,
    position = "center")
```

```
# creating spline variables
esoph$s1 <- esoph$age
esoph$s2 <- (esoph$age - 35) * (esoph$age > 35)
esoph$s3 <- (esoph$age - 45) * (esoph$age > 45)
esoph$s4 <- (esoph$age - 55) * (esoph$age > 55)
esoph$s5 <- (esoph$age - 65) * (esoph$age > 65)
esoph$s6 <- (esoph$age - 75) * (esoph$age > 75)

cider.lspline <- glm(case ~ bin_cider + s1 + s2 + s3 + s4 + s5 +
    s6, family = binomial, data = esoph)

# creating reduced model
```

```r
cider.lspline_red <- glm(case ~ s1 + s2 + s3 + s4 + s5 + s6,
    family = binomial, data = esoph)
# anova(cider.lspline, cider.lspline_red, test='LRT')

or <- exp(coef(cider.lspline)[2])
ci <- exp(confint.default(cider.lspline)[2, ])

kable(round(cbind(or, t(ci)), 2)) %>% kable_styling(full_width = F,
    position = "center")
```