

# BIOST 536: Homework 4

Department of Biostatistics @ University of Washington

October 25, 2024

Characteristic	Overall, N = 975	Cancer, N = 200	No cancer, N = 775
<b>Cider consumption, g/day</b>	11 (2, 31)	22 (4, 50)	10 (2, 26)
<b>Cider consumption, g/day</b>			
10 or less	477 (49%)	72 (36%)	405 (52%)
More than 10	498 (51%)	128 (64%)	370 (48%)
<b>Age, yr</b>	52 (41, 63)	61 (53, 67)	49 (39, 62)
<b>Age, yr</b>			
25-34	116 (12%)	1 (0.5%)	115 (15%)
35-44	199 (20%)	9 (4.5%)	190 (25%)
45-54	213 (22%)	46 (23%)	167 (22%)
55-64	242 (25%)	76 (38%)	166 (21%)
65-74	161 (17%)	55 (28%)	106 (14%)
75+	44 (4.5%)	13 (6.5%)	31 (4.0%)

1. Fit a logistic model relating binary cider exposure ( $> 10$  g/day) to the risk of esophageal cancer, using grouped-linear adjustment for age and the six age groups in the variable **agegp**. Report the odds ratio and 95% confidence interval for binary cider exposure. Optional: Write a sentence appropriate for the results section of a scientific paper reporting the results of this analysis. Note, here and below: your sentence should include both a point estimate and CI for the parameter(s) of interest.

Based on a logistic regression model adjusting for ordinal age levels (see Table 1), the odds of esophageal cancer for patients consuming over 10 grams of cider per day is 2.19 times (95% CI: 1.55-3.08; Robust Wald test  $p$ -value $<0.001$ ) the odds for patients of the same age level consuming 10 or less grams of cider per day. This provides evidence that daily cider consumption beneath 10 g offers health benefit over consumption above 10 g with adjustment of age level

2. Fit a logistic model relating binary cider exposure ( $> 10$  g/day) to the risk of esophageal cancer using indicator variables to adjust for age and the six age groups in the variable **agegp**. Report the odds ratio and 95% confidence interval for binary cider exposure. Optional: Write a sentence appropriate for the results section of a scientific paper reporting the results of this analysis.

Based on a logistic regression model adjusting for nominal age groups (see Table 1), the odds of esophageal cancer for patients consuming over 10 grams of cider per day is 2.03 times (95% CI: 1.44-2.85; Robust Wald test  $p$ -value $<0.001$ ) the odds for patients of the same age group consuming 10 or less grams of cider per day. This provides evidence that daily cider consumption beneath 10 grams offers health benefit over consumption of 10 or more grams with adjustment of age group.

- 3.

- (a) Compare the results for the exposure variable in the Q1 and Q2 analyses. Are results similar or very different?
  - (b) (Optional) Which result would you prefer to report in a scientific article, and why?
4. Fit a logistic model relating binary cider exposure ( $> 10$  g/day) to the risk of esophageal cancer using linear adjustment for age (as a continuous variable). Report the odds ratio and 95% confidence interval for binary cider exposure. Optional: Write a sentence appropriate for the results section of a scientific paper reporting the results of this analysis.

Based on a logistic regression model adjusting for age, the odds of esophageal cancer for patients consuming over 10 grams of cider per day is 2.17 times (95% CI: 1.54-3.06; Robust Wald test  $p$ -value $<0.001$ ) the odds for patients of the same age consuming 10 or less grams of cider per day. This provides evidence that daily cider consumption beneath 10 grams offers health benefit over consumption of 10 or more grams with adjustment of age group.

5. Fit a logistic model relating binary cider exposure ( $> 10$  g/day) to the risk of esophageal cancer using quadratic adjustment for age (as a continuous variable). Report the odds ratio and 95% confidence interval for binary cider exposure.

Based on a logistic regression model with quadratic adjustment for age, the odds of esophageal cancer for patients consuming over 10 grams of cider per day is 1.99 times (95% CI: 1.41-2.79; Robust Wald test  $p$ -value $<0.001$ ) the odds for patients of the same age consuming 10 or less grams of cider per day. This provides evidence that daily cider consumption beneath 10 grams offers health benefit over consumption of 10 or more grams with adjustment of age group.

6.

- (a) Fit a logistic model relating binary cider exposure ( $> 10$  g/day) to the risk of esophageal cancer using linear spline adjustment for age. Use the same age groups as for the `ageg` variable for your splines. Report the odds ratio and 95% confidence interval for binary cider exposure.

Based on a logistic regression model linear spline adjusting for age, the odds of esophageal cancer for patients consuming over 10 grams of cider per day is 1.99 times (95% CI: 1.41-2.79; Robust Wald test  $p$ -value $<0.001$ ) the odds for patients of the same age consuming 10 or less grams of cider per day. This provides evidence that daily cider consumption beneath 10 grams offers health benefit over consumption of 10 or more grams with linear spline adjustment of age.

- (b) (Optional) Write 1-2 sentences describing the analysis (not the results) that would be appropriate for the methods section of a scientific paper.

7.

- (a) Compare the OR estimates from Q4, Q5, and Q6. Are your results similar or very different?

OR associated with exposure, with linear adjustment for age: 2.17 OR associated with exposure, with quadratic adjustment for age : 1.99 OR associated with exposure, with linear spline adjustment for age : 1.98

- (b) (Optional) Which approach would you prefer if you were studying this exposure and wanted to adjust for age as a potential confounder, and why?

8. For each model in Q1 through Q6, how many degrees of freedom are used to include age in the model? That is, how many model parameters are used for age? Include the model intercept in your count.

Q1:  $\text{logit}(p) = \beta_0 + \beta_1 \text{Age}G$

Q2:  $\text{logit}(p) = \beta_0 + \beta_1 \text{Age}G_2 + \beta_2 \text{Age}G_3 + \beta_3 \text{Age}G_4 + \beta_4 \text{Age}G_5 + \beta_5 \text{Age}G_6$

Q4:  $\text{logit}(p) = \beta_0 + \beta_1 \text{Age}$

Q5:  $\text{logit}(p) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Age}^2 + \beta_3 \text{Age}^3 + \beta_4 \text{Age}^4$

Q6:  $\text{logit}(p) = \beta_0 + \beta_1 \text{Age} + \beta_2 (\text{Age} - 35)^+ + \beta_3 (\text{Age} - 45)^+ + \beta_4 (\text{Age} - 55)^+ + \beta_5 (\text{Age} - 65)^+ + \beta_6 (\text{Age} - 75)^+$

9. For each question (a)-(e), if you say that models are nested, write out the full and “reduced” models, and give the restrictions on the full model that yield the reduced model.

- (a) Are the models in Q1 and Q2 nested?
- (b) Are the models in Q1 and Q4 nested?
- (c) Are the models in Q4 and Q5 nested?
- (d) Are the models in Q4 and Q6 nested?
- (e) Are the models in Q5 and Q6 nested?

**End of report. Code appendix begins on the next page.**

## Code Appendix

```
# clear environment
rm(list=ls())

# setup options
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
options(knitr.kable.NA = '-', digits = 2)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "llm_appendix", "allcode")]
# load relevant packages
library(dplyr)      # data frame manipulation
library(ggplot2)
library(gtsummary)  # table summaries
# library(dagR)      # DAG simulation
library(broom)      # model coefficient table
# library(knitr)     # pretty tables
library(rigr)       # regression

# load data
load("../data/esophcts.Rdata")
esph <- esophcts
# names(esph)
# dim(esph)

# create variables for disease, exposure, and covariates
esph$bicider <- ifelse(esph$cider > 10, 1, 0) # cider consumption >10 g/day
esph$agegp <- as.factor(esph$agegp)
levels(esph$agegp) <- c("25-34", "35-44", "45-54", "55-64", "65-74", "75+")

## table summary of outcome, exposure, and covariates
esph %>%
  # relabel disease and exposure labels
  dplyr::mutate(
    D = ifelse(case==1, "Cancer", "No cancer"),
    E = cider,
    BiE = ifelse(bicider==1, "More than 10", "10 or less"),
    C = age,
    FaC = agegp,
    # only keep newly defined columns
    .keep="none") %>%
  # create table summary
  gtsummary::tbl_summary(
    by = D, label = list(c(BiE, E)~"Cider consumption, g/day",
                        c(C, FaC)~"Age, yr")) %>%
  add_overall() %>%
  bold_labels()

## =====
## Question 1
## =====

## fit logistic regression model of risk of esophageal cancer
```

```

mod <- rigr::regress("odds", case ~ bicider + as.numeric(agegp), esph)
coef <- coef(mod)[,c(4:6,8)]
# or, using glm()
# mod <- glm(case ~ bicider + as.numeric(agegp), family=binomial(), esph)
# coef <- cbind('e(Est)'=coef(mod), confint(mod)) %>% exp()
## =====
## Question 2
## =====

## fit logistic regression model of risk of esophageal cancer
mod2 <- regress("odds", case ~ bicider + agegp, esph)
coef2 <- coef(mod2)[,c(4:6,8)]

## =====
## Question 4
## =====

## fit logistic regression model of risk of esophageal cancer
mod3 <- regress("odds", case ~ bicider + age, esph)
coef3 <- coef(mod3)[,c(4:6,8)]

## =====
## Question 5
## =====

## fit logistic regression model of risk of esophageal cancer
mod4 <- regress("odds", case ~ bicider + poly(age,4), esph)
coef4 <- coef(mod4)[,c(4:6,8)]
# or, using glm()
# mod <- glm(case ~ bicider + I(age^4), family=binomial(), esph)
# coef <- cbind('e(Est)'=coef(mod), confint(mod)) %>% exp()

## =====
## Question 6 INCOMPLETE
## =====

## fit logistic regression model of risk of esophageal cancer
mod5 <- regress("odds", case ~ bicider + age + s1+s2+s3+s4+s5, esph)
coef5 <- coef(mod5)[,c(4:6,8)]

```

End of document.