BIOST 536: Homework 8

Department of Biostatistics @ University of Washington

4 December 2024

Background

In 1960-1961, 3,154 healthy, middle-aged men were entered into the Western Collaborative Group Study (WCGS), a long-term study of coronary heart disease (CHD). The risk of coronary heart disease mortality was studied for several variables measured at baseline, i.e., Type A/B behavior, systolic blood pressure, serum cholesterol level, cigarette smoking status, and age. Although there is no codebook for the WCGS dataset, the variable names in the "wcgsdata.Rdata" dataset are largely informative. As done in Lecture 9, we will pretend that the variable `chd69` is a scientifically valid binary indicator of coronary heart disease, ignoring the important issue of differential length of follow-up.

Logistic Regression

1. Use logistic regression to investigate the question: Does smoking status modify the association between systolic blood pressure (SBP) and CHD. Fit an appropriate logistic regression model that addresses this question using a binary smoking exposure indicator variable (0 if ncigs==0, 1 if ncigs>0).

a. Write your fitted logistic regression model.

$$logit(p) = -6.578 + 0.029 \times SBP + 0.973 \times SMOKE + -0.002 \times SMOKE \times SBP$$

b. Explain briefly how to use the model to test the null hypothesis that smoking does not modify the SBP-CHD association.

A likelihood ratio test of $\beta_3 = 0$ will determine if exclusion of effect modification significantly improves the modeled likelihood of the data.

c. Perform an appropriate statistical test corresponding to b. Write the results in a sentence suitable for a scientific publication.

We fit a logistic regression of the log odds of CHD from SBP with effect modification from smoking status. A likelihood ratio test determined smoking status does not moderate the association of CHD with SBP, or vice versa (LRT p = .75). A similar test using the Wald test statistic with robust standard errors produced the same p-value.

d. Regardless of the results in (c), use your fitted model to describe the SBP association with CHD for smokers and non-smokers and include confidence intervals.

We estimate the odds ratio (OR) associated with a 1-unit increase in SBP to be 1.00 among nonsmokers (95% CI for OR among nonsmokers: 1.00-1.02) and among smokers (95% CI for OR among smokers: 1.00-1.00). This is an odd result.

Poisson Regression

2.

a. Use a regression technique that summarizes associations with relative risks. Investigate an association between systolic blood pressure (SBP) and having a CHD event. Write out the fitted model, then write 1-2 sentences summarizing results in language suitable for a scientific publication.

$log(p) = log(P(CHD = 1)) = -5.584 + 0.023 \times SBP$

We fit a modified Poisson regression of the log risk of CHD from SBP. According to the model, the relative risk (RR) associated with a 1-unit increase in SBP is 1.024 (95% robust CI for RR: 1.017-1.03).

b. (part b is optional) In class we discussed three different variants of relative risk regression. Use a different variant than the one you used for 3a and compare/contrast results.

We failed to fit a log-binomial model or non-linear least squares model, due to numerical instability. Instead, we fit a linear regression of the risk of CHD from SBP. According to the model, the risk difference (RD) associated with a 10-unit increase in SBP is 0.024 (95% robust CI for RD: 0.018-0.031).

3. Does being a smoker modify the association between systolic blood pressure (sbp) and CHD as measured with the relative risk? Fit an appropriate model that addresses this question using a binary smoking exposure variable.

a. Write your fitted model.

$log(p) = -6.1829 + 0.0253 \times SBP + 0.9681 \times SMOKE - 0.0027 \times SBP \times SMOKE$

b. Perform an appropriate statistical test to answer the question. Write the results in a sentence suitable for a scientific publication.

We fit a modified Poisson regression of the log risk of CHD from SBP with effect modification from smoking status. A robust Wald test determined smoking status does not moderate the association of CHD with SBP (Wald p = .68).

4. Comment on differences or similarities between results in Q1 and Q3.

Both logistic and modified Poisson regression found insufficient evidence of meaningful interaction between smoking status and SBP on the association with CHD risk. The two models produced similar estimates, although the main effect of SBP is estimated to be greater by logistic regression.

Linear Regression

5. Now use a regression technique that summarizes associations with risk differences. Investigate an association between systolic blood pressure (SBP) and having a CHD event, adjusting for age, body mass index (BMI) and cholesterol (chol). Write out the fitted model, then write 1-2 sentences summarizing results in language suitable for a scientific publication.

$log(p) = \beta_0 + \beta_1 SBP + \beta_2 age + \beta_3 bmi + \beta_4 chol$

Weight is avaible in the WCGS data set, not BMI; we will use weight as a substitute.

We fit a linear regression of the risk of CHD from SBP, adjusting for age, weight, and cholesterol. According to the model, the risk difference associated with a 10-unit increase in SBP is 0.016 (95% robust CI for RD: 0.010 - 0.023).

Poisson Rate Regression

6. We will now revisit the Framingham data and consider modeling CHD incidence and its association with high cholesterol. We will use the "framingham_HW8.Rdata" dataset. Note: Incidence=#chd events/person-time at risk. Assume there are 365.25 days in a year use R to generate the variable logpyears = log(days/365.25). Also, create a binary CHDbin variable which is 0 if CHD="Censored" and 1 if CHD="CHD".

a) Use Poisson regression to estimate the association between high cholesterol (chol200=1) and incidence of CHD, adjusted for sex, age (agegrp) and body mass index (BMIgrp). State your conclusion in a sentence or two suitable for an abstract in a medical journal.

We fit a Poisson regression of the log rate of CHD incidence from high cholesterol status, adjusting for sex, age group, and BMI group. According to the model, the incidence rate ratio (IRR) associated with high cholesterol is 1.513 (95% robust CI for IRR: 1.328-1.723).

b) From the model in 6a, estimate the incidence and 95% CI of CHD (per 1,000 person-years) for males over 60 with BMI > 30 and cholesterol over 200, and then calculate the incidence for the same group but with cholesterol below or equal to 200. For this dataset, the model and robust variance estimates are similar, so you can use the estimable command to get good approximations of the confidence intervals you need.

According to the previous model, the incidence of CHD (per 1,000 person-years) of the first group is 60 (95% CI: 49-75). For the second group, our estimate is 40 (95% CI: 31-51)

7. Fit a slightly different model compared to Q1. Use a different coding for the binary smoking variable (-1 if ncigs==0, 1 if ncigs>0). With this coding, fit a model with a main effect for sbp, main effect for smoking (coded -1/1), and an interaction term.

a. Explain how to use this model to test the null hypothesis that smoking does not modify the SBP association with CHD.

This null hypothesis is the same as before and will indicate the same conclusion.

b. Compare the p-value for the interaction term from this model to the p-value for the interaction term for Q1. Are they the same or different? Why does this make sense?

I have yet to fit it, but I believe they should be the same.

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, warning=FALSE, message=FALSE, results='hide')
options(knitr.kable.NA = '-', digits = 2)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "allcode")]
# Load relevant packages
library(gmodels)  # create model estimates with estimable()
library(dplyr)    # data manipulation
library(lmtest)   # testing linear regression models

# Load WCGS data
wcgs <- read.csv("../data/wcgs.csv")
# Load FRAMINGHAM data
load("../data/framingham_HW8.Rdata")
framingham <- fdata

## Handle missing data
anyNA(wcgs)  # TRUE
colMeans(is.na(wcgs)) %>% subset(. > 0)  # negligible missingness
wcgs <- na.omit(wcgs)

anyNA(framingham)  # TRUE
colMeans(is.na(framingham)) %>% subset(. > 0)  # negligible missingness
framingham <- na.omit(framingham)

## Data processing
wcgs <- rename(wcgs, chd = chd69)  # rename CHD variable for convenience

## Final inspection
head(wcgs)
head(framingham)
#### --- QUESTION 1 --- ####

# Create an indicator for smoking status
wcgs$smoke <- ifelse(wcgs$ncigs == 0, 0, 1)

## Fit a logistic regression of log odds of CHD from SBP with interaction from smoking status
fit1 <- glm(chd ~ sbp * smoke, family = binomial, wcgs)
# Conduct a LRT
anova(fit1, test = "LRT")[4,5]  # LRT p = .75
# Compare to a Wald test with robust SE
lmtest::coeftest(fit1, cvov = "sandwich")[4,4]  # Robust Wald p = .75

# Get estimated SBP effect and CI for smokers
gmodels::estimable(fit1, c("sbp"=1), conf=0.95)[c(1,6,7)] %>% exp
# Get estimated SBP effect and CI for nonsmokers
estimable(fit1, c("sbp"=1, "sbp:smoke"=1), conf=0.95)[c(1,6,7)] %>% exp
```

```
#### --- QUESTION 2 --- ####

## Fit a Poisson regression of log risk of CHD from SBP
fit2 <- glm(chd ~ sbp, family = poisson(log), wcgs)
# Get coefficient estimates and robust CI
fit2_rawcoef <- cbind(Estimate=coef(summary(fit2))[,1], coefci(fit2, cvov="sandwich"))
exp(fit2_rawcoef)

## Fit a log-binomial regression of log risk of CHD from SBP
# glm(chd ~ sbp, family = gaussian(identity), wcgs)  # FAILS

## Fit a Gaussian regression (non-linear least squares) of log risk of CHD from SBP
# glm(chd ~ sbp, family = gaussian(identity), wcgs)  # FAILS

## Fit a Gaussian regression of risk of CHD from SBP
fit3 <- lm(chd ~ sbp, wcgs)
# Get coefficient estimates and robust CI
fit3_rawcoef <- cbind(Estimate=coef(summary(fit3))[,1], coefci(fit3, cvov="sandwich"))
10*fit3_rawcoef

#### --- QUESTION 3 --- ####

## Fit a Poisson regression of log risk of CHD from SBP with interaction from smoking status
fit4 <- glm(chd ~ sbp * smoke, family = poisson(log), wcgs)
# Get coefficient estimates and robust CI
fit4_rawcoef <- cbind(Estimate=coef(summary(fit4))[,1], coefci(fit4, cvov="sandwich"))
exp(fit4_rawcoef)
# Conduct a Wald test with robust SE
coeftest(fit4, cvov = "sandwich")[4,4]  # Robust Wald p = .68

#### --- QUESTION 5 --- ####

## Fit a linear regression of risk of CHD from SBP
fit5 <- lm(chd ~ sbp + age + weight + chol, wcgs)
# Get raw coefficient estimates with robust SE
fit5_rawcoef <- cbind(Estimate=coef(summary(fit5))[,1], coefci(fit5, cvov="sandwich"))
fit5_rawcoef*10

#### --- QUESTION 6 --- ####

# Create indicator for observation of a CHD event
framingham$chdbin <- ifelse(framingham$chd == "CHD", 1, 0)

## Fit a Poisson regression of rate of CHD incidence from SBP, adjusting for sex, age group, and BMI gr
fit6 <- glm(chdbin ~ chol200 + sex + agegrp + bmigrp, offset = log(days/365.25),
            family = poisson(log), framingham)
# Get raw coefficient estimates with robust SE
fit6_rawcoef <- cbind(Estimate=coef(summary(fit6))[,1], coefci(fit6, cvov="sandwich"))
exp(fit6_rawcoef)

# Get estimated incidence and CI for first group
1000*exp(estimable(fit6,
                   c("(Intercept)"=1, "agegrp60+"=1, "bmigrp>30"=1, "chol200"=1),
```

```
                  conf=0.95))[c(1,6,7)]
# Get estimated incidence and CI for second group
1000*exp(estimable(fit6,
                  c("(Intercept)"=1, "agegrp60+"=1, "bmigrp>30"=1),
                  conf=0.95))[c(1,6,7)]
# Create alternative indicator for smoking status
wcgs$smoke2 <- ifelse(wcgs$ncigs == 0, -1, 1)
```

End of document.