Apply multiple linear regression to study the association between all-cause mortality and creatinine, age, smoking history, sex, and race

Eliza Chai

02/22/2021

Responses

1. The fitted logistic model is as follows:

$\hat{log(odds}(Y_i))$ = -9.248 + 1.378 $*$ $creatinine_i$ + 0.081 $*$ age_i + 0.409 $*$ 1_{smoker_i} + 0.302 $*$ 1_{sex_i} - 0.381 $*$ $1_{race:white_i}$ - 0.130 $*$ $1_{race:black_i}$ - 0.107 $*$ $1_{race:asian_i}$.

$\hat{log(odds}(Y_i))$ is the fitted log-odds where Y_i is an indicator of subject i died during MRI study. $creatinine_i$ is the creatinine level of participant i measured in mg/dl. age_i is the age of participant i in years. 1_{smoker_i} = 1 if participant i is a smoker, $1_{smoker_i} = 0$ otherwise. $1_{sex_i} = 1$ if the sex of participant i is male, 1_{sex_i} = 0 if the sex of participant i is female. $1_{race:white_i} = 1$ if the race of participant i is White, $1_{race:white_i} = 0$ otherwise. $1_{race:black_i} = 1$ if the race of participant i is Black, $1_{race:black_i} = 0$ otherwise. $1_{race:asian_i} = 1$ if the race of participant i is Asian, $1_{race:asian_i} = 0$ otherwise.

2. Summarize the findings about the association between all-cause mortality and creatinine, age, smoking history, sex, and race:

Method:

Data were obtained from the MRI cohort to study the association between all-cause mortality and creatinine, age, smoking history, sex, and race. With the MRI dataset, we fit a multiple logistic regression model with log odds of all-cause mortality as the response and creatinine, age, smoking history, sex, and race as predictors. In the fitted multiple logistic regression model, we applied heteroskedasticity-robust standard errors and performed tests at a 5% significance level.

Results:

We estimated that the odds ratio for two populations who differ by one mg/dl in the creatinine level but have the same age, smoking history, sex, and race is 3.966 (95% CI based on robust standard errors: 1.993, 7.889), with the higher creatinine group having higher odds of death. We conclude that there is significant association between creatinine level and mortality after adjusting for race, age, sex and smoking history ($p < 0.0005$).

We estimated that the odds ratio for two populations who differ by one year in age but have the same creatinine level, smoking history, sex, and race is 1.085 (95% CI based on robust standard errors: 1.046, 1.125), with the older group having higher odds of death. We also find a significant association between age and all-cause mortality after adjusting for race, creatinine level, sex and smoking history ($p < 0.0005$).

For subpopulations of the same creatinine level, age, sex, and race but who differ in their smoking history, we estimated that the odds of all-cause mortality during the study is 1.505 times greater for a population who

smokes compared to the non-smoking population (95% CI based on robust standard error: 0.981 - 2.308). We find no statistically significant association between smoking history and odds of death at the 5% level (p = 0.061) after adjusting for age, creatinine level, race and sex.

In addition, the odds of all-cause mortality during the study is estimated to be 1.352 times greater for a male population compared to a female population of the same creatinine level, age, smoking history, and race (95% CI based on robust standard error: 0.878 - 2.081). We also do not find a statistically significant association between sex and all-cause mortality after adjusting for age, creatinine level, race and smoking history (p = 0.170).

For subpopulations of the same creatinine level, age, smoking history, and race but who differ in race, the odds of all-cause mortality during the study is estimated to be 0.683 times greater for a White population compared to a non-White/Black/Asian population (95% CI based on robust standard error: 0.163 - 2.862). We find no statistically significant of association between race being White and all-cause mortality after adjusting for age, creatinine level, sex and smoking history (p = 0.602). The odds of all-cause mortality are estimated to be 0.878 times greater for a Black population compared to a non-White/Black/Asian population of the same creatinine level, age, sex, and smoking history (95% CI based on robust standard error: 0.196 - 3.933). We find no statistically significant of association between race being Black and all-cause mortality after adjusting for age, creatinine level, sex and smoking history (p = 0.865). Furthermore, the odds of all-cause mortality are estimated to be 0.899 times greater for an Asian population compared to a non-White/Black/Asian population of the same creatinine level, age, sex, and smoking history (95% CI based on robust standard error: 0.181 - 4.467). We also find no statistically significant of association between race being Asian and all-cause mortality after adjusting for age, creatinine level, sex and smoking history. Overall, we do not find a significant association between race and all-cause mortality after adjustment for creatinine level, age, sex and smoking history (LRT p = 0.71).

Thus, we conclude that we have evidence of significant associations between all-cause mortality with both creatinine level and age, but not smoking history, sex, and race (controlling for the other variables for each association).

3. From the results above, there is a statistically significant first-order linear trend ($p < 0.0005$) between all-cause mortality and creatinine level. In addition, there is a statistically significant first-order linear trend ($p < 0.0005$) between all-cause mortality and age. Therefore, we conclude that we have evidence of first-order linear associations between all-cause mortality with both creatinine level and age in the positive direction, which is not a surprising association to observe. Since a high level of creatinine build-up in the blood is taken as an indication of kidney disease and older in age are more likely to have worsened body functions, we expect these predictors to be positively associated with all-cause mortality. However, from the results above, there is no statistically significant first-order linear trend between all-cause mortality with smoking history (p = 0.061). It is surprising that we find no evidence of first-order linear associations between smoking history and all-cause mortality as smoking behavior are causally associated with decreased lung functions and worsened clinical outcome. For sex as a predictor, there is no statistically significant linear trend at the 5% level (p = 0.170). It is not surprising that we find no evidence of linear associations between sex and all-cause mortality since we would not expect an association between sex and mortality in general. For race as a predictor, there is no statistically significant linear trend at the 5% level. It is also not surprising that we find no evidence of first-order linear associations between race and all-cause mortality since we would not expect an association between race and mortality.

4. The sample odds ratio for disease given exposure is $\frac{d/c}{b/a} = \frac{ad}{bc}$ The sample odds ratio for exposure given disease is $\frac{d/b}{c/a} = \frac{ad}{bc}$

Number of participants	Y=0	Y=1
X=0	a	b

Number of participants	Y=0	Y=1
X=1	c	d

5. Use Bayes' Theorem to show that the true odds ratio for disease given exposure equals the true odds ratio for exposure given disease.

$$\text{OR} = \frac{odds(Y=1|X=1)}{odds(Y=1|X=0)} = \frac{P(Y=1|X=1)/[1-P(Y=1/X=1)]}{P(Y=1|X=0)/[1-P(Y=1/X=0)]} = \frac{odds(X=1|Y=1)}{odds(X=1|Y=0)} = \frac{(P(X=1|Y=1)/[1-P(X=1/Y=1)])}{P(X=1|Y=0)/[1-P(X=1/Y=0)]} = \frac{ad}{bc}$$

Code Appendix

```
### Setting up the packages, options we'll need:
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(rigr)
### ------------------------------------------------------------
### Reading in the data.
mri <- read_csv("mri.csv")
### ------------------------------------------------------------
### Q1
### Here's a logistic model for mri dataset using rigr:regress
### the response variable is death and predictor variables are creatinine,
### age, smoking status, sex, and race
mri <- mri %>%
  mutate(smoke_smoker = ifelse(packyrs > 1, 1, 0)) %>%   #create indicator of smoking status:smoker
  mutate(sex_male = ifelse(sex == "Male", 1, 0)) %>%   #create indicator of sex:male
  mutate(race_white = ifelse(race == "White", 1, 0)) %>%   #create indicator of race:white
  mutate(race_black = ifelse(race == "Black", 1, 0)) %>%   #create indicator of race:black
  mutate(race_asian = ifelse(race =="Asian", 1, 0))   #create indicator of race:asian

mod1 <- regress("odds", death ~ crt + age + smoke_smoker + sex_male +
                  race_white + race_black + race_asian, data = mri)
mod1 %>% coef %>% round(3)
### ------------------------------------------------------------
### Q2
### Here's a logistic model for mri dataset using rigr:regress
### the response variable is death and predictor variables are creatinine,
### age, smoking status, sex, and race
mod1 %>% coef %>% round(3)
### Fit logistic model for mri dataset using glm
### Full model response variable death and predictor variables of
### creatinine, age, smoking status, sex, and race
mod_full <- glm(death ~ crt + age + smoke_smoker + sex_male + race_white +
                race_black + race_asian, data = mri, family = "binomial")
### Null model response variable death and predictor variables of
### creatinine, age, smoking status, and sex
mod_null <- glm(death ~ crt + age + smoke_smoker + sex_male, data = mri,
                family = "binomial")
### Run ANOVA test
anova(mod_null, mod_full, test = "LRT")
### ------------------------------------------------------------
### Q4
### Table of Disease (Y) & Exposure (X)
### Y=1 indicate diseases, X=1 indicates exposure
tab1 <- tibble(y=c("X=0", "X=1"),y1=c("a","c"), y2=c("b","d"))
names(tab1) <- c("Number of participants", "Y=0", "Y=1")
knitr::kable(tab1)
### ------------------------------------------------------------
```