BIOST 536: Homework 6

Department of Biostatistics @ University of Washington

15 November 2024

Background

The Studies of Left Ventricular Dysfunction (SOLVD) trials were two large randomized placebo-controlled trials of the efficacy of enalapril, an angiotensin converting enzyme (ACE) inhibitor, for the treatment and prevention of congestive heart failure to improve survival in patients with a weak left ventricular ejection fraction [1]. The original analyses were based on a time-to-event outcome; however, for the purposes of this homework, we consider a simulated version of the data that was based on this study but uses a binary version of the death outcome (`I_death`, 1=yes, 0=no) that ignores the differential length of follow-up. Covariates in this dataset include the treatment arm indicator (`I_ace`, 1=ACE Inhibitor, 0=Placebo control) and the baseline characteristics: sex (`I_sex`), age in years (`age`), left ejection fraction (`ef`), and an indicator of whether they had diabetes (`I_dm`). Please use the **sovld_simulated.csv** data set to answer the following set of questions.

REFERENCE [1] SOLVD Investigators. Studies of left ventricular dysfunction (SOLVD)—rationale, design and methods: two trials that evaluate the effect of enalapril in patients with reduced ejection fraction. The American Journal of Cardiology. 1990 Aug 1;66(3):315-22.

Data

We have sufficient sample sizes to model odds of death with logistic regression.

Assignment

1. Use logistic regression to assess the unadjusted effect of the ACE inhibitor in this trial on survival (i.e. on the death outcome). Provide the odds ratio (OR) and confidence interval.

We estimate the odds of death are 0.84 times lower for patients with a weak left ventricular ejection fraction treated with the ACE inhibitor compared to patients of the same population treated with a placebo control (95% CI for odds ratio: 0.74-0.95). Using a Rao test we conclude that there is a significant difference in the odds of death when comparing these two populations (p=0.0045).

2. Suppose in planning the study, the investigators hypothesized that age, sex, ejection fraction and diabetes were all potential precision variables. What would need to be true about the association with outcome and exposure for these variables to be precision variables? Does the data seem to support the idea that each of these variables might be useful as precision variables? State your reasoning.

Precision variables, introduced by linear regression, are variables that (1) are not directly associated with the predictor of interest, (2) causally affect the outcome, and (3) **???**. They may reduce a model's standard error and provide narrower confidence intervals. This concept extends to logistic regression for randomized trials, although an important caveat is that inclusion of a precision variable does not guarantee reduced standard errors. However, they generally increase effect size estimates and model power.

We can evaluate age, sex, ejection fraction and diabetes as potential precision variables by testing if they satisfy these criteria, both scientifically and statistically.

- **Age.** The distribution of age for subgroups of both death status and treatment appear to be right-skewed, although approximately normal. Using a Welch Two Sample t-test we conclude that there is a significant difference in mean age when comparing the two populations of death status ($p<0.001$). Also using a Welch Two Sample t-test we conclude that there is a significant difference in mean age when comparing the two treatment populations ($p<0.001$). The data does not suggest age may be a useful precision variable because it is associated with treatment.
- **Ejection fraction.** The distribution of ejection fraction for subgroups of both death status and treatment appear to be right-skewed, non-normal. Using a Wilcoxon rank sum test we conclude that there is a significant difference in mean ejection fraction when comparing the two populations of death status ($p<0.001$). Also using a Wilcoxon rank sum test we conclude that there is a significant difference in mean ejection fraction when comparing the two treatment populations ($p<0.001$). Again, the data does not suggest ejection fraction may be a useful precision variable because it is associated with treatment.
- **Patient's sex.** Using a Pearson's Chi-squared test with simulated p-value (based on 2000 replicates) we conclude that there is not a significant difference in proportion of females/males when comparing the two populations of death status ($p=0.5$). Using a similar test, we conclude that there is not a significant difference in proportion of females/males when comparing the two treatment populations ($p=0.3$). The data does not suggest a patient's sex may be a useful precision variable because it is not associated with the outcome, death status.
- **Diabetic status.** Using a Pearson's Chi-squared test with simulated p-value (based on 2000 replicates) we conclude that there is a significant difference in proportion of females/males when comparing the two populations of death status ($p<0.001$). Using a similar test, we conclude that there is not a significant difference in proportion of females/males when comparing the two treatment populations ($p=0.2$). The data suggests a patient's diabetic status may be a useful precision variable because it is associated with death status and not associated with treatment.

3. Fit the model in 1, now adding in age, sex, ejection fraction and diabetes. Did the association between treatment and outcome appear to get stronger, weaker or stay the same compared to the unadjusted model in #1?

Compared to the unadjusted model, the effect size is more strongly negative, aligning with our expectation that precision variables in logistic regression increase power and are not ensured to reduce standard error.

4. Fit a model to evaluate whether ejection fraction (as a continuous variable) modifies the treatment effect on survival. What do you conclude from this model and summarize your conclusion in a sentence suitable for a scientific publication.

We estimate the odds ratio associated with ACE inhibitor treatment to be 1.02 times greater for a patient population that differs from another patient population in ejection fraction by 1-unit, with the higher ejection fraction patients having higher odds ratio associated with treatment (95% CI for ratio of odds ratio: 1-1.04). Using a likelihood ratio test we conclude that ejection fraction (as a continuous variable) modifies the treatment effect on survival ($p=0.043$).

5. Now suppose the investigators wanted to evaluate whether having low ejection fraction (indicated by ef<20, yes/no) modifies the treatment's effect on death. Fit a model to evaluate this question and state your conclusion, including your rationale.

We estimate the odds ratio associated with ACE inhibitor treatment to be 0.72 times greater for a patient population with low ejection fraction (<20 units) compared to a patient population without low ejection fraction (>=20 units) (95% CI for ratio of odds ratio: 0.52-0.98). Using a likelihood ratio test we conclude that low ejection fraction (as a binary variable) modifies the treatment effect on survival (p=0.039).

6. Using the model in #5, provide the treatment odds ratio and 95% confidence interval for those with and without low ejection fraction (assuming constant values for any other variables that are in your model for this comparison).

We estimate the odds ratio associated with ACE inhibitor treatment to be 0.87 for a patient population without low ejection fraction (>=20 units) (95% CI for odds ratio: 0.76-1). For a patient population with low ejection fraction (<20 units), we estimate that the odds ratio associated with treatment is 0.62 (95% CI for odds ratio: 0.47-0.83)

7. Suppose the effect of the baseline ejection fraction on the death outcome is now of interest. Can you use the coefficients in the model you fit in question 5 to evaluate whether treatment modifies the effect of low ejection fraction on death? If yes, discuss what is your conclusion.

Yes, we may. Our estimate of the ratio of odds ratio associated with treatment comparing low ejection fraction patients to non-low ejection fraction patients is the same as the ratio of odds ratio associated with low ejection fraction comparing patients who received treatment to those who received placebo. Using the same likelihood ratio test result from before, we conclude that treatment modifies the effect of low ejection fraction (as a binary variable) on death (p=0.039).

8. (Optional) Investigate whether the modification of the treatment effect on the ejection fraction appears to be non-linear.

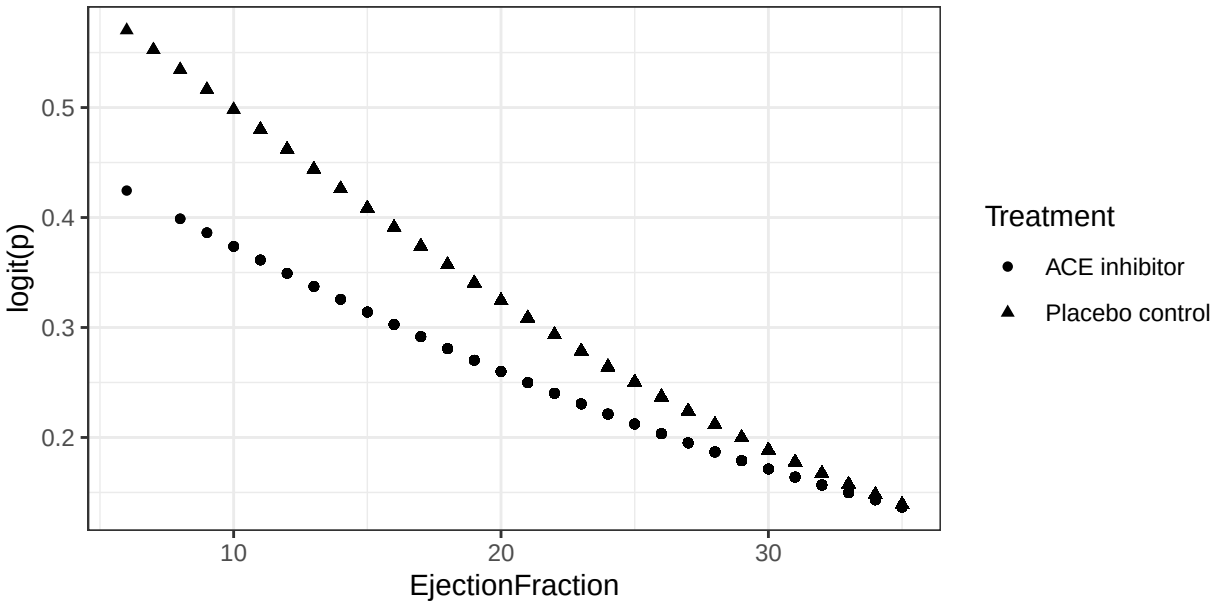

Figure 1: Log odds of death over ejection fractions, stratified by treatment

Plotting the predicted log odds of death across measures of ejection fraction, stratified by treatment group (Figure 1) suggests that the modification of ejection fraction on treatment effect may be non-linear.

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(dplyr)      # data manipulation
library(ggplot2)    # data visualization
library(gtsummary)  # data summary
library(broom)      # tidy tables
library(gmodels)    # model fitting

# load data
svd <- read.csv("../data/solvd_simulated.csv")
ls(); names(svd); dim(svd)
str(svd)

## handle missing data
anyNA(svd)  #no missing data

## create labelled data
svd_labelled <- svd %>%
  dplyr::mutate(
    Outcome = ifelse(I_death == 1, "Died", "Survived"),
    Treatment = ifelse(I_ace == 1, "ACE inhibitor", "Placebo control"),
    Sex = ifelse(I_sex == 1, "Male", "Female"),
    Age = age,
    Diabetes = ifelse(I_dm == 1, "Diabetic", "Not Diabetic"),
    EjectionFraction = ef,
    .keep = "none")

## create Table 1 summary
gtsummary::tbl_summary(svd_labelled, by = Outcome) %>%
  add_overall()

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

## model odds of death with logistic regression
mod1 <- glm(I_death ~ I_ace, family = binomial, data = svd)
coef1 <- broom::tidy(mod1, conf.int = TRUE, exponentiate = TRUE) %>%
  dplyr::rename(OR = estimate)

## evaluate model
coef1[2, c(2,7,6,5)]    # OR associated with ACE inhibitor treatment
anova(mod1, test="Rao")   # Rao (aka Score) test for significance of treatment
```

```r
## ===============
## Question 2
## ===============

#### === AGE === ####

## association with Death
# descriptive statistics
svd_labelled %>% group_by(Outcome) %>%
  summarize(Mean = mean(Age), SD = sd(Age),
            Median = median(Age), IQR = IQR(Age))
# histogram
ggplot(svd_labelled, aes(Age)) + geom_histogram() + facet_grid(vars(Outcome))
# t-test works because approx. normal
with(svd, t.test(I_death, I_sex))

## association with Treatment
# descriptive statistics
svd_labelled %>% group_by(Treatment) %>%
  summarize(Mean = mean(Age), SD = sd(Age),
            Median = median(Age), IQR = IQR(Age))
# histogram
ggplot(svd_labelled, aes(Age)) + geom_histogram() + facet_grid(vars(Treatment))
# t-test works because approx. normal
with(svd, t.test(I_ace, I_sex))

## t-test for association with death, p-value: <2e-16
## t-test for association with treatment, p-value: <2e-16

#### === EJECTION FRACTION === ####

## association with Death
# frequency
svd_labelled %>% group_by(Outcome) %>%
  summarize(Mean = mean(EjectionFraction), SD = sd(EjectionFraction),
            Median = median(EjectionFraction), IQR = IQR(EjectionFraction))
# histogram
ggplot(svd_labelled, aes(EjectionFraction)) + geom_histogram() +
  facet_grid(vars(Outcome))
# t-test won't work because not approx. normal; use Wilcoxon
with(svd, wilcox.test(I_death, ef))

## association with Treatment
# frequency
svd_labelled %>% group_by(Treatment) %>%
  summarize(Mean = mean(EjectionFraction), SD = sd(EjectionFraction),
            Median = median(EjectionFraction), IQR = IQR(EjectionFraction))
# histogram
ggplot(svd_labelled, aes(EjectionFraction)) + geom_histogram() +
  facet_grid(vars(Treatment))
# t-test won't work because not approx. normal; use Wilcoxon
with(svd, wilcox.test(I_ace, ef))
```

```r
## Wilcoxon test for association with death, p-value: <2e-16
## Wilcoxon test for association with treatment, p-value: <2e-16

#### === SEX === ####

## association with Death
with(svd_labelled, table(Outcome, Sex) / nrow(svd_labelled))
with(svd, chisq.test(I_death, I_sex, simulate.p.value = TRUE))

## association with Treatment
with(svd_labelled, table(Treatment, Sex) / nrow(svd_labelled))
with(svd, chisq.test(I_ace, I_sex, simulate.p.value = TRUE))

## chi-squared test for association with death, p-value: 0.5
## chi-squared test for association with treatment, p-value: 0.3

#### === DIABETES === ####

## association with Death
with(svd_labelled, table(Outcome, Diabetes) / nrow(svd_labelled))
with(svd, chisq.test(I_death, I_dm, simulate.p.value = TRUE))

## association with Treatment
with(svd_labelled, table(Treatment, Diabetes) / nrow(svd_labelled))
with(svd, chisq.test(I_ace, I_dm, simulate.p.value = TRUE))

## chi-squared test for association with death, p-value: 0.0005
## chi-squared test for association with treatment, p-value: 0.2

## ===============
## Question 3
## ===============

# model odds of death
formula2 <- I_death ~ I_ace + age + I_sex + ef + I_dm
mod2 <- glm(formula2, family = binomial, data = svd)
coef2 <- tidy(mod2, conf.int=TRUE, exponentiate=TRUE) %>% rename(OR = estimate)

# get estimated OR associated with ACE inhibitor
cbind(model=c(1,2), rbind(coef1[2, c(2,5:7)], coef2[2, c(2,5:7)]))

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

# model odds of death
mod3 <- glm(I_death ~ I_ace + ef, family = binomial, data = svd)
mod4 <- glm(I_death ~ I_ace * ef, family = binomial, data = svd)
coef3 <- tidy(mod3, conf.int=TRUE, exponentiate=TRUE) %>% rename(OR = estimate)
coef4 <- tidy(mod4, conf.int=TRUE, exponentiate=TRUE) %>% rename(OR = estimate)
coef3; coef4

# test if interaction term significantly improve model likelihood
```

```r
anova(mod3, mod4, test="LRT")  # LRT p-value: 0.043

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

# create low ejection fraction variable
svd$efLow = ifelse(svd$ef < 20, 1, 0)
# svd_labelled$EFBinary = ifelse(svd$efLow == 1, "Low", "Not low")

# model odds of death
mod5 <- glm(I_death ~ I_ace + efLow, family = binomial, data = svd)
mod6 <- glm(I_death ~ I_ace * efLow, family = binomial, data = svd)
coef5 <- tidy(mod5, conf.int=TRUE, exponentiate=TRUE) %>% rename(OR = estimate)
coef6 <- tidy(mod6, conf.int=TRUE, exponentiate=TRUE) %>% rename(OR = estimate)
coef5; coef6

# test if interaction term significantly improve model likelihood
anova(mod5, mod6, test="LRT")  # LRT p-value: 0.039

# create models for two populations of `efLow`, validate with their coefficients
glm(I_death ~ I_ace, family = binomial, data = subset(svd, efLow==0)) %>%
  tidy(exponentiate=TRUE, conf.int=TRUE) %>% rename(OR = estimate)
glm(I_death ~ I_ace, family = binomial, data = subset(svd, efLow==1)) %>%
  tidy(exponentiate=TRUE, conf.int=TRUE) %>% rename(OR = estimate)

## ===============
## Question 8
## ===============

# plot logit(p) across EF
svd_labelled$mod4pred <- mod4$fitted.values

ggplot(svd_labelled, aes(x=EjectionFraction, y=mod4pred, shape=Treatment)) +
  geom_point() + ylab("logit(p)") + theme_bw()
```

End of document.