PB HLTH 250C: Assignment 2

Due Thursday 29 February 2024 at 11:59pm via Gradescope

Read all questions carefully before answering. You may work in small groups of no more than 3 individuals and turn in a single assignment (and everyone in the group will receive the same grade). Work through the entire assignment individually first, then come together to discuss and collaborate. Please maintain numbering on sub-questions (if any), type your responses, and **please keep answers brief.** Also, start each answer on a new page, and make sure to **link the questions on Gradescope to the corresponding page in your PDF document.**

Bootstrap confidence interval for standardized estimates

We will use boostrapping to estimate confidence intervals for estimates of association for risk of death comparing two *populations* that only differ in their distribution of blood glucose.[1] Elevated glucose is a risk factor for many serious diseases, including heart disease, stroke, and cancer. Glucose levels are often classified into 3 categories according to the American Diabetes association (ADA) criterion.[2]

Preliminary tasks

First read in the dataset `Glucose_data.Rdata` that contains the variables:

- `death`: a binary indicator of mortality status after 20 years of follow-up.
- `glucose.cat`: categorized blood glucose levels, according to ADA criterion:
  - <100 mg/dL (Normal blood glucose).
  - 100- <126 mg/dL (Prediabetes).
  - >=126 mg/dL (Diabetes).
- `sex` is a dichotomous variable indicating male/female sex.
- `educ` is a categorical variable indicating educational level:
  - 1=less than high-school education
  - 2=high school graduate
  - 3=some college
  - 4=college graduate
- `cursmoke` is a dichotomous variable indicating if the individual was a current smoker.

```r
library(dplyr) # for recoding variable
library(boot)
load("Glucose_data.Rdata") # Make sure to set working directory or add pathname
```

[1] Revisiting part of the PH 252 assignment on regression standardization.

[2] For refresher on model-based standardization, refer to the Generalized Linear Models lecture in PB HLTH 252, and the paper by Greenland, S "Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies." *American Journal of Epidemiology*. 160.4 (2004): 301-305.

Create a function to calculate the standardized RR and RD

On your own, complete the following function that:

1. Fits a logistic regression model on a resampled dataset.
2. Creates datasets of pseudo-populations corresponding to the scenarios of interest.
3. Calculates and returns the population-averaged risk differences and standardized risk ratios for the scenarios of interest.

```r
standardized.measures <- function(dataset, index){

    # Create resampled version of dataset using index vector:
    data.resamp <-  ##### COMPLETE THIS STEP

    fit.logistic <- glm(death ~ factor(glucose.cat) + age + factor(sex) +
                            factor(educ) + factor(cursmoke),
                        data=data.resamp,
                        family=binomial)

    ### To calculate the measures of association:

    # STEP 1: Create two new versions of the *original* dataset, called
    #         data.resamp.low (the pseudo-population "everyone with lowest glucose category")
    #         data.resamp.red (the pseudo-population "everyone reduces glucose category 1 level")

    data.resamp.low <- data.resamp.red <- ## COMPLETE THIS STEP

    ### Set blood glucose categories in comparison datasets:

    # Pseudo-sample if everyone was in lowest category:
    data.resamp.low$glucose.cat <- "[0,100)"

    # Pseudo-sample if those w/ elevated glucose reduced their levels
    # (lowered by 1 category):
    data.resamp.red$glucose.cat <- recode(data.resamp.red$glucose.cat,
                                    "[100,126)"="[0,100)",
                                    "[126,Inf)"="[100,126)")

    # STEP 2: Obtain predicted individual risk of hypertension under each new dataset:
    rhat.obs <- predict(fit.logistic, type="response") # uses data from model fit
    rhat.low <- # COMPLETE THIS STEP
    rhat.red <- # COMPLETE THIS STEP

    # STEP 3: Calculate the average risk (proportion) of death
    #         in each hypothetical population:
    risk.obs <- mean(rhat.obs)
    risk.low <- mean(rhat.low)
    risk.red <- mean(rhat.red)

    ### Calculate risk differences and risk ratios
    ##  A. Everyone with low glucose levels vs. the observed distribution:
    rd.low.obs <- ##### COMPLETE THIS STEP
    rr.low.obs <- ##### COMPLETE THIS STEP

    ##  B. Reduced glucose levels vs. the observed distribution:
    rd.red.obs <- ##### COMPLETE THIS STEP
    rr.red.obs <- ##### COMPLETE THIS STEP


    # STEP 4: Return these estimates:
    return(c(rd.low.obs, rr.low.obs,
             rd.red.obs, rr.red.obs))
}
```

Obtain point estimates and confidence intervals

Call the function you defined, using the index values for the original dataset to obtain the point estimates for the standardized *RD* and *RR*:

```
n.obs <- nrow(glucose.data)
stdized.measures <- ### COMPLETE THIS STEP

stdized.rd <- stdized.measures[1] # RD low vs. obs
stdized.rr <- stdized.measures[2] # RD reduced vs. obs
```

Use the boot package to obtain 95% bias-corrected and accelerated confidence intervals for these measures (this may take a while to run):[3]

```
set.seed(123)
# Put the bootstrapped sample results into object called bs.standardized
bs.standardized <- boot(glucose.data, standardized.measures, R=5000,
                        parallel="multicore", ncpus=4)
# Summarize each of the 4 bootstrap samples:
boot.ci(bs.standardized, type= "bca", index=1) # rd.low.obs
boot.ci(bs.standardized, type= "bca", index=2) # rr.low.obs
boot.ci(bs.standardized, type= "bca", index=3) # rd.red.obs
boot.ci(bs.standardized, type= "bca", index=4) # rr.red.obs
```

Plot the estimated density from the bootstrapped samples for both sets of results (use a single multi-panel plot for each pair). Note that the series of samples is in the element t which has two columns (first corresponds to the RD and second to the RR):

```
##### GRAPH LOW GLUCOSE VS. OBSERVED DISTRIBUTION
rd.samples <- bs.standardized$t[,1]
rr.samples <- bs.standardized$t[,2]

par(mfrow=c(2,1))
plot(density(rd.samples), main="Bootstraped Samples of Risk Difference",
     xlab="RD",
     sub="Low glucose vs. observed distribution")
plot(density(rr.samples), main="Bootstraped Samples of Risk Ratio",
     xlab="RR",
```

[3] If the parallel processing option doesn't work for you, let us know and we can try to help troubleshoot, and/or make a note on your submission.

```
     sub="Low glucose vs. observed distribution")
par(mfrow=c(1,1))

##### GRAPH REDUCED GLUCOSE VS. OBSERVED DISTRIBUTION
###  COMPLETE THIS STEP
```

On your own:

1. Calculate the standard deviation of the series of bootstrapped samples for the RD and RR **for the low vs. observed glucose distribution scenarios** obtained in the above code chunk. **Calculate these centered on the original point estimate (not the sample means).**
2. Use these estimates of standard deviation to calculate 95% confidence intervals for the RD and RR **for the low vs. observed comparisons** using a Normal approximation approach.

Questions

Bootstrap confidence interval (assume all CIs are two-sided)

For reporting risk difference (RD) measures ONLY, multiply point estimates and confidence intervals by 1,000, report to 2 decimal places, and interpret effect per 1,000 individuals. (NOT for relative measures.)

1. Present the point estimate and 95% bias-corrected and accelerated (BCa) CI for all 4 of the estimated standardized risk difference and standardized risk ratios for this analysis. **Briefly** (no more than a few sentences) interpret the point and interval estimates as you would in the results of a manuscript. (**20 points**)

2. Turn in the two sets of density plots that you produced for the risk difference and risk ratio measures for the bootstrapped samples. **Make sure they are correctly and clearly labeled.** Briefly describe the pattern that you see in the distribution of each (skewness, shape of the tails, location of the mode, etc. . .). Does the pattern seem to indicate that the sampling distribution of the RD and RR are approximately normal? (**10 points**)

3. **Showing your work**, calculate the 95% CI for the risk difference (RD) **for the low vs. observed glucose distribution** comparison using a Normal approximation, and report your answer.[4] What assumptions does this require with regard to the distribution of the estimator of the RD? Based on the plots from Q2 do you think this assumption was fulfilled? How does this CI compare to the BCa result from question 1? (**10 points**)

4. Showing your work, calculate the 95% CI for the risk ratio (RR) **for the low vs. observed glucose distribution** comparison, and report your answer.[3] What assumptions does this require with regard to the distribution of the estimator of the RR? Based on the plots from Q2 do you think this assumption was fulfilled? How does this CI compare to the BCa result from question 1? (**10 points**)

[4] Show the formula you used, and the specific values that went into your calculation.