01_question1

44

2022-06-02

Q1 A trial was performed to assess the impact of an intervention on the number of times a sedentary individual engages in 10 minutes of sustained moderate-to-vigorous physically active over a week.

The study recruited participants in matched pairs (matching on age), assigning the intervention to one individual within each age pair (n total pairs).

Denote each pair using the subscript i and let the number of active bouts for the person in the ith pair be denoted as Y_{i1}, Y_{i2}, with Y_{i1} corresponding to individual who received the intervention.

Let X_i be the age for individuals in the ith pair.

It is hypothesized that

$$Y_{i1}|X_i \sim Poisson\big(\gamma\beta(X_i)\big)$$
$$Y_{i2}|X_i \sim Poisson\big(\beta(X_i)\big)$$

where $\beta(\cdot)$ is an unspecified function of X_i (age).

The investigators of the current study are uninterested in $\beta(\cdot)$.

Assume that the distribution of age in the population follows some distribution $f_X(x)$.

If the number of pairs (n) observed is large and X is observed densely relative to the shape of $\beta(\cdot)$ then γ, β may be estimable using the likelihood of the data.

However, if n is small or X is sparsely observed, estimating both γ and β may not be possible. This question involves deriving a conditional likelihood where you condition out the nuisance parameter(s) $\beta(\cdot)$ so that γ is estimable in such scenarios.

You will implement a short simulation study and apply your method to data from the study referenced above.

(a) In the first part of this problem, we will condition out nuisance parameters (β) to obtain a likelihood for γ which depends only on the observed data.

i. What is the likelihood of $[X_i, Y_{i1}, Y_{i2}]^T|\gamma, \beta(\cdot)$? Your answer will involve the marginal distribution of X, $f_X(x)$.

$$
\begin{aligned}
L(\gamma, \beta(\cdot); \mathbf{X}, \mathbf{Y}_1, \mathbf{Y}_2) &= \prod_{i=1}^{n} L\left(\gamma, \beta(\cdot); X_i, Y_{i1}, Y_{i2}\right) \\
&= \prod_{i=1}^{n} f_{Y_1,Y_2}\left(y_{i1}, y_{i2}|x_i; \gamma, \beta(\cdot)\right) \times f_X\left(x_i; \gamma, \beta(.)\right) \\
&= \prod_{i=1}^{n} f_{Y_1}\left(y_{i1}|x_i; \gamma, \beta(.)\right) \times f_{Y_2}\left(y_2|x_i; \beta(.)\right) \times f_X(x_i) \\
&= \prod_{i=1}^{n} \frac{(\gamma\beta(x_i))^{y_{i1}} e^{-\gamma\beta(x_i)}}{y_{i1}!} \frac{(\beta(x_i))^{y_{i2}} e^{-\beta(x_i)}}{y_{i2}!} f_X(x_i) \\
&= \prod_{i=1}^{n} \frac{\gamma^{y_{i1}} \beta(x_i)^{y_{i1}+y_{i2}} e^{-\left(\gamma\beta(x_i)+\beta(x_i)\right)}}{y_{i1}! y_{i2}!} f_X(x_i)
\end{aligned}
$$

ii. What is the likelihood of $Y_{i1} + Y_{i2}|Xi, \theta$? Identify the corresponding distribution.

The summation of two Poisson distribution is also a Poisson distribution. $Y_{i1} + Y_{i2}|Xi \sim Poisson\big(\gamma\beta(X_i) + \beta(X_i)\big)$

$$\begin{cases} U = Y_1 + Y_2 \\ V = Y_1 \end{cases}$$
$$Y_1, Y_2 \geq 0; 0 \leq V \leq U$$
$$\begin{cases} Y_1 = V \\ Y_2 = U - V \end{cases}$$
$$J = \begin{bmatrix} \frac{\partial u}{\partial y_1} & \frac{\partial u}{\partial y_2} \\ \frac{\partial v}{\partial y_1} & \frac{\partial v}{\partial y_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} = -1$$

$$\begin{aligned} f_{U,V}\big(u, v|x; \gamma, \beta(.)\big) &= f_{Y_1,Y_2}(y_1 + y_2, y_1|x; \gamma, \beta(.)) \times |J| \\ &= \frac{\gamma^v \beta(x)^u e^{-(\gamma\beta(x)+\beta(x))}}{v!(u-v)!} \\ &for:\ 0 \leq V \leq U_i \\ f_U\big(u|x; \gamma, \beta(.)\big) &= \int f_{U,V}\big(u, v|x; \gamma, \beta(.)\big)dv \\ &= \int_0^u \frac{\gamma^v \beta(x_i)^{u_i} e^{-\left(\gamma\beta(x_i)+\beta(x_i)\right)}}{v_i!(u_i-v_i)!} dv \\ &= \frac{\left(\gamma\beta(x_i)+\beta(x_i)\right)^{u_i} e^{-\gamma\beta(x_i)-\beta(x_i)}}{u_i!} \\ U_i|X_i \equiv Y_{i1} + Y_{i2}|X_i &\sim Poisson\big(\gamma\beta(X_i) + \beta(X_i)\big) \end{aligned}$$

iii. What is the likelihood of $Y_{i1}|Y_{i1}+Y_{i2}, X_i, \theta$? Identify the corresponding distribution.

$$
\begin{aligned}
L(\theta; v|u,x) &= f_{V|U}(v|u,x;\theta) = \frac{f_{U,V}(v,u|x;\theta)}{f_U(u|x;\theta)} \\
&= \frac{\gamma^v \beta(x_i)^{u_i} e^{-(\gamma\beta(x_i)+\beta(x_i))}}{v_i!\,(u_i-v_i)!} \times \frac{u_i!}{\left(\gamma\beta(x_i)+\beta(x_i)\right)^{u_i} e^{-\gamma\beta(x_i)-\beta(x_i)}} \\
&= \frac{u_i!}{v_i!\,(u_i-v_i)!}\frac{\gamma^{v_i}\beta^{u_i}}{(\gamma\beta+\beta)^{u_i}} \\
&= \frac{u_i!}{v_i!\,(u_i-v_i)!}\left(\frac{\gamma\beta}{\gamma\beta+\beta}\right)^{v_i}\left(\frac{\beta}{\gamma\beta+\beta}\right)^{u_i-v_i} \\
&= \frac{u_i!}{v_i!\,(u_i-v_i)!}\left(1-\frac{1}{\gamma+1}\right)^{v_i}\left(\frac{1}{\gamma+1}\right)^{u_i-v_i}
\end{aligned}
$$

The distribution of $(V_i|U_i, X_i;\theta) \equiv (Y_{i1}|Y_{i1}+Y_{i2}, X_i;\theta)$ is $(V_i|U_i, X_i;\theta) \sim Binomial\left(U_i, \frac{1}{\gamma+1}\right)$, with sample size U_i and $p = \frac{1}{\gamma+1}$

The distribution for the parameter θ is $\frac{1}{1+\gamma} \sim Beta(u_i - v_i + 1, v_i + 1)$

iv. What is the maximum likelihood estimator of γ obtained from the conditional likelihood you derived in iii?

If $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$; $\gamma = g(\theta) = \theta^{-1} - 1$

$$L(\gamma; \mathbf{v}|\mathbf{u}, \mathbf{x}) = \prod_{i=1}^{n} L(\theta; v_i|u_i, x_i)$$

$$\log L(\gamma; \mathbf{v}|\mathbf{u}, \mathbf{x}) = \sum_{i=1}^{n} \log L(\theta; v_i|u_i, x_i)$$

$$= \sum_{i=1}^{n} \log\left(\frac{u_i!}{v_i!\,(u_i - v_i)!}\right) + \log\left(1 - \frac{1}{\gamma+1}\right)\sum_{i=1}^{n} v_i + \log\left(\frac{1}{\gamma+1}\right)\sum_{i=1}^{n}(u_i - v_i)$$

$$\frac{\partial \log L(\gamma; \mathbf{v}|\mathbf{u}, \mathbf{x})}{\partial \gamma} = \frac{1}{\gamma(\gamma+1)}\sum_{i=1}^{n} v_i - \frac{1}{(\gamma+1)}\sum_{i=1}^{n}(u_i - v_i) \overset{set}{=} 0$$

$$\hat{\gamma} \neq -1 \text{ or } 0$$

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} v_i}{\sum_{i=1}^{n}(u_i - v_i)} = \frac{\sum_{i=1}^{n} y_{i1}}{\sum_{i=1}^{n} y_{i2}} = \frac{\bar{Y}_1}{\bar{Y}_2}$$

$$\frac{\partial^2 \log L(\gamma; \mathbf{v}|\mathbf{u}, \mathbf{x})}{\partial \gamma^2} = -\frac{2\gamma+1}{\gamma^2(\gamma+1)^2}\sum_{i=1}^{n} v_i + \frac{1}{(\gamma+1)^2}\sum_{i=1}^{n}(u_i - v_i)$$

$$given\ \hat{\gamma},\quad \frac{\partial^2 \log L(\gamma; \mathbf{v}|\mathbf{u}, \mathbf{x})}{\partial \gamma^2} \leq 0;\ \hat{\gamma}\ is\ MLE$$

(b) Propose a method for constructing an asymptotic 95% confidence interval for γ based on your result from (a) iv using maximum likelihood theory (you may use the observed information here).

The MLE $\hat{\gamma}$ is asymptotically normal under regularity conditions.

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{L} Normal(0, I(\gamma)^{-1})$$

$$Var[\hat{\gamma}] = I[\gamma]^{-1} = -\left(\frac{1}{n}\frac{\partial^2}{\partial \gamma^2}\log L(\gamma; \mathbf{v}|\mathbf{u}, \mathbf{x})\right)^{-1} = \frac{(\bar{Y}_1 + \bar{Y}_2)\,(\bar{Y}_1)}{\bar{Y}_2^3}$$

Hence, an asymptotic 95% confidence interval for γ can be calculated using $\hat{\gamma} \pm 1.96 se[\hat{\gamma}]$

(c) A statistician colleague of yours suggests constructing of a confidence interval using simulation. Specifically, your colleague proposes the following procedure:

Fix α, the type-I error rate.

- Step 1 Choose candidate interval endpoints $\gamma^* = [\gamma_1^*, \gamma_2^*, \ldots, \gamma_S^*]^\top$ which is dense over a range of plausible endpoints for the confidence interval.
- Step 2 For each $\gamma_s^* \in \gamma^*$
  - Step 2a Simulate $y_{i1}^s | y_{i1} + y_{i2}, \gamma_s^*$ using the distribution you derived in (a)iii.
  - Step 2b Calculate γ_s^* as the MLE you derived in (a)iv.
  - Step 2c Repeat (1)-(2) many times (e.g. 2000).
  - Step 2d Let $I_s = 1\left(\hat{\gamma} \in \left[\hat{F}_{\gamma_s}^{-1}(\alpha/2), \hat{F}_{\gamma_s}^{-1}(1-\alpha/2)\right]\right)$, where $\hat{F}_{\gamma_s}^{-1}(\alpha/2))$ and $\hat{F}_{\gamma_s}^{-1}(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of γ_s^* obtained from Steps 2a-2c, respectively.
- Step 3 Define a $1-\alpha\%$ CI for γ

Argue that your colleague's algorithm is a valid approach for constructing a 95% confidence interval. Write a function which implements this algorithm for arbitrary: - Type-I error α - Input vectors $\mathbf{y}_1, \mathbf{y}_2$ - Candidate interval endpoints γ^* - Number of simulated datasets created in step 2c

We are trying to find with a random γ_s^* whether we can find the interval covers the true MLE value. It is the same as using a stochastic process to find the empirical distribution of γ. According to Donsker's theroem and Glivenko-Cantelli theorem, we know that the empirical process distribution convergence to the true distribution asymptotically. Hence, we can use sampling to generate a valid confidence interval.

$$(V_i | U_i, X_i; \theta) \sim Binomial\left(U_i, \frac{1}{\gamma+1}\right)$$

$$\hat{\gamma} = \frac{\sum_{i=1}^n y_{i1}}{\sum_{i=1}^n y_{i2}} = \frac{\bar{Y}_1}{\bar{Y}_2}$$

```
library(tidyverse)
## simulation function.
gety <- function(seed,
                 n = 15) {
  set.seed(seed)
  X <- runif(n, min = 0, max = 1)
  beta <- 2 * exp(-2 * X)
  gamma <- 2.2

  y1 <- map(gamma * beta, ~rpois(1, .)) %>% unlist()
  y2 <- map(beta, ~rpois(1, .)) %>% unlist()

  simy <- cbind(y1 = y1, y2 = y2)
```

```r
  return(simy)
}

## simulate 5000 y1 y2 set.
sim_num <- 5000
y_ss15 <- map(1:sim_num, ~gety(seed = ., n = 15))
y_ss100 <- map(1:sim_num, ~gety(seed = ., n = 100))
save(y_ss15, y_ss100, file = "bios_qexam_data_1.Rdata")

load("bios_qexam_data_1.Rdata")

y1 <- map(y_ss15, ~mean(.[, 1])) %>% unlist()
y2 <- map(y_ss15, ~mean(.[, 2])) %>% unlist()
mean(y1); mean(y2); mean(y1) / mean(y2); sqrt((mean(y1) + mean(y2))^3 * mean(
y2) / (mean(y1)^3 * 15))
mean(y1) / mean(y2) + 1.96 * sqrt((mean(y1) + mean(y2))^3 * mean(y2) / (mean(
y1)^3 * 15))
mean(y1) / mean(y2) - 1.96 * sqrt((mean(y1) + mean(y2))^3 * mean(y2) / (mean(
y1)^3 * 15))
power.t.test(n = 15, delta = 2.2 -1, sd = 0.421, sig.level = 0.05)

y1 <- map(y_ss100, ~mean(.[, 1])) %>% unlist()
y2 <- map(y_ss100, ~mean(.[, 2])) %>% unlist()
mean(y1); mean(y2); mean(y1) / mean(y2); sqrt((mean(y1) + mean(y2))^3 * mean(
y2) / (mean(y1)^3 * 100))
mean(y1) / mean(y2) + 1.96 * sqrt((mean(y1) + mean(y2))^3 * mean(y2) / (mean(
y1)^3 * 100))
mean(y1) / mean(y2) - 1.96 * sqrt((mean(y1) + mean(y2))^3 * mean(y2) / (mean(
y1)^3 * 100))
power.t.test(n = 100, delta = 1.2, sd = 0.163, sig.level = 0.05)

load("bios_qexam_data_1.Rdata")

## function for question 1d
get_gamma <- function(y,
                      alpha = 0.05,
                      gamma_min = 0,
                      gamma_max = 5.5,
                      gamma_step = 0.01,
                      N = 2000) {
  ## sum of y1 y2
  y1 <- y[, 1]
  y2 <- y[, 2]
  y12 <- y1 + y2
  gamma_int <- seq(gamma_min, gamma_max, by = gamma_step)
  gamma_hat <- sum(y1) / sum(y2)
  ## emptry matrix to store the values
  gamma_mle <- matrix(NA, nrow = N, ncol = length(gamma_int))

  for (j in seq_along(1:N)) {
```

```r
    for (i in seq_along(1:length(gamma_int))) {
      phat <- gamma_int[i] / (1 + gamma_int[i])
      ## simulate y1 | y1+y2 2000 times
      # ya3 <- map(y12, ~rbinom(n = 1, size = .,
      #                         prob = phat)) %>% unlist()
      ya3 <- rbinom(nrow(y1),
                    size = unlist(y1 + y2),
                    prob = phat)
      gamma_mle[j, i] <- sum(ya3) / sum(y12 - ya3)
    }
  }

  ## get F(0.25) and F(0.975)
  q025 <- apply(gamma_mle, 2,
                quantile, probs = alpha / 2,
                na.rm = TRUE)

  q975 <- apply(gamma_mle, 2,
                quantile, probs = 1 - alpha / 2,
                na.rm = TRUE)
  ## check whether gamma_int is located inside F025 F975
  gamma_ind <- ifelse((gamma_hat > q025) & (gamma_hat < q975), 1, 0)
  ## find the smallest and largest gamma_int
  lower <- min(gamma_int[which(gamma_ind == 1)])
  upper <- max(gamma_int[which(gamma_ind == 1)])
  return(c(lower, upper))
}

# get_gamma(y_ss100[[1]])

## parallel programming
library(furrr)
library(parallel)
detectCores()

## use 6 cores
plan(multisession, workers = 6)
finall_5k_N15 <- future_map_dfc(y_ss15, ~get_gamma(y = .))
finall_5k_N100 <- future_map_dfc(y_ss100, ~get_gamma(y = .))

save(finall_5k_N15, finall_5k_N100, finall_per,
     file = "bios_qexam_question1.Rdata")

load("bios_qexam_question1.Rdata")
library(matrixStats)
rowMeans(as.matrix(finall_5k_N15))

# get power

sum(ifelse(1 > finall_5k_N15[1,] & 1 < final_5k_N15[2, ], 1, 0)) / 5000
```

```
sum(ifelse(1 > finall_5k_N100[1,] & 1 < final_5k_N100[2, ], 1, 0))/ 5000

rowMeans(as.matrix(finall_5k_N100))
```

(d) Implement a fully reproducible simulation study which assesses the performance of the confidence intervals obtained in (b) and (c). Specifically, implement a simulation study as follows:

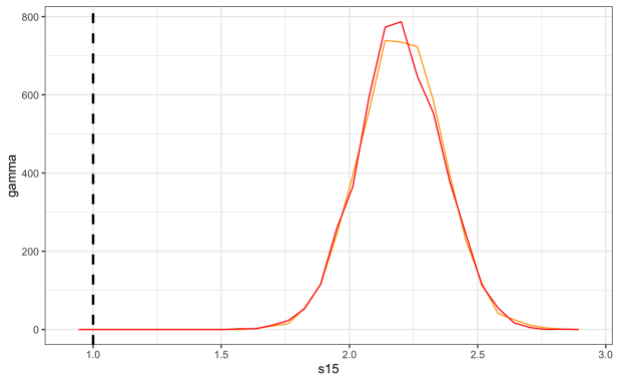
- $X \sim Unif(0,1)$,
- $\gamma = 2.2$,
- $\beta(x) = 2e^{-2x}$,
- $n = 15 \& n = 100$,
- γ^* is an equally spaced grid on [0,5.5] with interval length of 0.01
- Simulate 2000 datasets in Step 2c of the algorithm from (c)

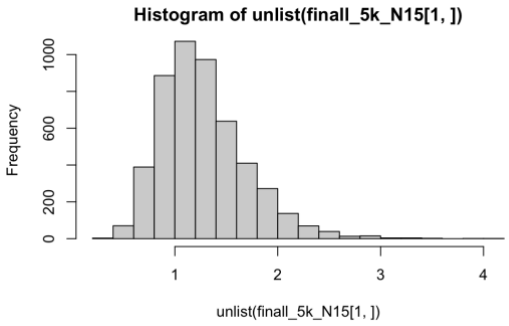
Estimate 95% confidence intervals using the approaches you derived in both (b) and (c) in 5000 simulated datasets and evaluate power for testing the null hypothesis $H_0: \gamma = 1$. Compare approaches on coverage probability and power. Summarize your results in 3-4 sentences as if for a statistical journal. Support your claims using 1-2 tables or figures. Provide well commented code which reproduces your simulation study and recreates your tables/figures.

The confidence interval calculated with asymptotic features 1b: sample size 15 MLE with mean 2.200, sd 0.421, and 95% confidence interval [1.375, 3.025]; sample size 100 MLE with mean 2.199, sd 0.163, and 95% confidence interval [1.879, 2.518]

This final 95% confidence interval does not include value of 1. However, the coverage rate is 0.259, hence the power is 0.741 for sample size 15; coverage rate is 0.005, the power is 0.995. As we can see in the graph below, the location of $\gamma = 1$ is far away from the asymptotic distribution of MLE (not included in the 95% CI), which means we have little change to fail reject the null hypothesis.

The confidence interval based on method in 1d, 1 is not included in the 95% confidence interval in any simulation, the power should be larger than 0.95 in all the cases near to 1. Hence, we can reject the null hypothesis. The confidence interval for sample size 15 is [1.790, 3.115]; The confidence interval for sample size 100 is [1.952, 2.973]





(e) Apply your methods to the dataset "MVPA intervention.csv", a data matrix which contains two columns, "Y1" and "Y2", corresponding to the number of 10 minute bouts of moderate-to-vigorous physical activity in matched pairs. Estimate γ and construct 95% confidence intervals using both of the approaches you derived in (b) and (c).

Write a brief, non-technical, summary of both methods as if communicating to a non-statistician, recommend and justify one approach, and conclude with an interpretation of the estimated coefficient and confidence interval.

The asymptotically normal distribution of MLE with mean 3.5 and sd 1.403, which provides the 95% confidence interval [0.749, 6.250]. This result is based on asymptotic theory with large sample size, which might be biased and misleading with small sample size.

The empirical method in 1d provides the confidence interval (1.71, 7.56). To confirm this result, a permutation test with 1000 resampling was applied. This provides similar wider confidence interval [1.825, 8.64].

With small sample size, I would not claim either method is perfect. Especially the second empirical method is computationally very intense. However, I will recommend the second empirical methods. Because the sample size as 15 is small, that we cannot get a precise confident interval with asymptotic assumptions. Especially the confidence interval is not much larger for the second method. Hence, we can conclude that the mean value for γ is 3.5, and if we randomly generate confidence interval, there is at least 95% of the chance that [1.71, 7.56] will cover the true γ value. In less accurate terms, we are 95% confident that the population parameter is between [1.71, 7.56]. The later interpretation will mislead the reader to regard the true parameter value as random.

```
library(tidyverse)
y <- read_csv("data/MVPA_intervention.csv")
y1 <- y$Y1
y2 <- y$Y2
y1_bar <- mean(y1)
y2_bar <- mean(y2)
## mle
(mean <- y1_bar / y2_bar)
## sd
## use observed information
(gamma <- y1_bar / y2_bar)
Inof <- ((2 * gamma + 1) / (gamma^2 * (1 + gamma)^2) * sum(y1) - 1 / (gamma +
 1)^2 * sum(y2)) / 15
```

```r
(sd <- (1 / Inof) %>% sqrt()) / sqrt(15); sd
## use phat
sd0 <- sqrt((y1_bar + y2_bar) * y1_bar / y2_bar^3) / sqrt(15)
## 95% CI
mean - 1.96 * sd0
mean + 1.96 * sd0

get_gamma(y = y, gamma_max = 15)

## use permutation to get the confidence interval
y_permu <- map(1:1000, ~sample_n(y, size = 15, replace = TRUE))

## parallel programming
library(furrr)
library(parallel)

## use 6 cores
plan(multisession, workers = 6)
finall_per <- future_map_dfc(y_permu, ~get_gamma(y = ., gamma_max = 10))
rowMedians(as.matrix(finall_per))
## [1] 1.67 9.95
## [1] 1.825 8.640
```