# Question 2

Joseph Froelicher

3/12/2021

## Part A

We need two parameters to simulate a logistic regression to model the outcome using the only the exposure, and those are $\beta_0$ and $\beta_1$. In order to calculate $\beta_0$ and $\beta_1$, we need the probability of disease, given exposure $P(Y = 1|X = 1)$, and the probability of disease, given no exposure $P(Y = 1|X = 0)$.

$$\beta_0 = ln\left(\frac{P(Y = 1|X = 0)}{1 - P(Y = 1|X = 0)}\right)$$

$$\beta_1 = ln\left(\frac{\left[\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)}\right]}{\left[\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)}\right]}\right)$$

$\beta_0$ = -2.944439, $\beta_1$ = 0.7472144

## Part B

```
# Part B
set.seed(8675309)

x = rbinom(n, s, pe)
pee = plogis(b0 + b1 * x)
y = rbinom(n, s, pee)

data = data.frame(
  "id" = 1:n,
  "outcome" = y,
  "exposure" = x
)

# check population proportions
sume = 0
counte = 0
sumu = 0
countu = 0

for (i in 1:dim(data)[1]) {
  if (data$exposure[i] == 0) {
    sumu = sumu + data$outcome[i]
    countu = countu + 1

  } else {
    sume = sume + data$outcome[i]
```

```
    counte = counte + 1
  }
}

head(data)
```

```
##   id outcome exposure
## 1  1       0        0
## 2  2       0        0
## 3  3       0        1
## 4  4       0        1
## 5  5       0        0
## 6  6       0        0
```

```
tail(data)
```

```
##            id outcome exposure
## 99995   99995       0        0
## 99996   99996       0        1
## 99997   99997       0        1
## 99998   99998       0        1
## 99999   99999       0        0
## 100000 100000       1        1
```

```
# test proportion unexposed
sumu / countu
```

```
## [1] 0.05028004
```

```
#test proportion exposed
sume / counte
```

```
## [1] 0.09958184
```

## Part C

```
head(estimates)
```

```
##   seed     b0_hat        b1_hat
## 1    1 -20.566069  1.875078e+01
## 2    2  -3.178054  7.357068e-01
## 3    3  -1.658228 -1.093307e+00
## 4    4  -2.751535  1.005090e-15
## 5    5  -2.751535  3.091883e-01
## 6    6  -3.178054 -1.359740e-15
```

```
tail(estimates)
```

```
##      seed    b0_hat        b1_hat
## 995   995 -3.891820 -2.410623e-14
## 996   996 -2.751535  3.091883e-01
## 997   997 -2.197225  6.808771e-01
## 998   998 -2.197225  8.109302e-01
## 999   999 -2.751535  1.093307e+00
## 1000 1000 -3.891820  2.375473e+00
```

The mean of $\beta_0$ across all simulations is -4.330406 and the median is -3.1780538. The mean of $\beta_1$ across all simulations is 1.8929855 and median is 0.7591051. These are not quite where we would want them to be

compared to part A, because we are taking a fairly small sample size ($n = 50$). As we increase the sample size, and number of iterations in the simulation, we will converge on our calculated estimates of $\beta_0$ and $\beta_1$ from part A.

### Part D

```
head(estimatesd)
```

```
##   seed     b0_hat    b1_hat
## 1    1 -0.5306283 1.1592369
## 2    2 -0.1251631 0.3483067
## 3    3 -0.4187103 1.1526795
## 4    4 -0.1177830 0.3690975
## 5    5 -0.3930426 1.2259517
## 6    6 -0.3920421 1.0459686
```

```
tail(estimatesd)
```

```
##      seed     b0_hat    b1_hat
## 995   995 -0.1698990 0.4150215
## 996   996 -0.4054651 1.1856237
## 997   997 -0.4054651 1.4170660
## 998   998 -0.2363888 0.7472144
## 999   999 -0.2006707 0.5029516
## 1000 1000 -0.3254224 0.8644189
```

The mean of $\beta_0$ across all simulations is -0.2718071 and the median is -0.268264. The mean of $\beta_1$ across all simulations is 0.7464198 and median is 0.7357068.

### Part E

The median absolute deviance is lower for the case-control study (0.4517607) than the cohort study (0.8949844). This is indicating the the efficiency of the log odds ratio may be better in the case-control style study, where we are modeling exposure based on disease, rather than modeling disease based on exposures. This is consistent with our understanding of case-control studies and cohort studies. Cohort studies require much larger populations, and case-control studies reach high power at lower population sizes.

## Appendix

```
library(tidyverse)

# Part A
n = 100000
s = 1
pde = 0.1
pdu = 0.05
pe = 0.3

b0 = log(pdu / (1 - pdu))
b1 = log(pde / (1 - pde) / (pdu / (1 - pdu)))

# Part B
set.seed(8675309)
```

```r
x = rbinom(n, s, pe)
pee = plogis(b0 + b1 * x)
y = rbinom(n, s, pee)

data = data.frame(
  "id" = 1:n,
  "outcome" = y,
  "exposure" = x
)

# check population proportions
sume = 0
counte = 0
sumu = 0
countu = 0

for (i in 1:dim(data)[1]) {
  if (data$exposure[i] == 0) {
    sumu = sumu + data$outcome[i]
    countu = countu + 1

  } else {
    sume = sume + data$outcome[i]
    counte = counte + 1
  }
}

head(data)
tail(data)

# test proportion unexposed
sumu / countu

#test proportion exposed
sume / counte

# Part C
# cohort study
N = 50
b = 1000

datae = data[data$exposure == 1,]
datau = data[data$exposure == 0,]
estimates = data.frame(
  "seed" = rep(NA, b),
  "b0_hat" = rep(NA, b),
  "b1_hat" = rep(NA, b)
)

for (i in 1:b) {
  set.seed(i)
  exposed = sample(datae$outcome, N )
  unexposed = sample(datau$outcome, N)
```

```
  temp_dat = data.frame(
    "exposure" = c(rep(1, N), rep(0, N)),
    "outcome" = c(exposed, unexposed)
  )

  model = glm(outcome ~ exposure, family = binomial, data = temp_dat)
  estimates$seed[i] = i
  estimates$b0_hat[i] = summary(model)$coefficients[1, 1]
  estimates$b1_hat[i] = summary(model)$coefficients[2, 1]
}

mean_b0 = mean(estimates$b0_hat)
mean_b1 = mean(estimates$b1_hat)
median_b0 = median(estimates$b0_hat)
median_b1 = median(estimates$b1_hat)

head(estimates)
tail(estimates)
# Part D
# case-control study
datad = data[data$outcome == 1,]
datan = data[data$outcome == 0,]
estimatesd = data.frame(
  "seed" = rep(NA, b),
  "b0_hat" = rep(NA, b),
  "b1_hat" = rep(NA, b)
)

for (i in 1:b) {
  set.seed(i)
  disease = sample(datad$exposure, N)
  nodisease = sample(datan$exposure, N)
  temp_dat = data.frame(
    "outcome" = c(rep(1, N), rep(0, N)),
    "disease" = c(disease, nodisease)
  )

  model = glm(outcome ~ disease, family = binomial, data = temp_dat)
  estimatesd$seed[i] = i
  estimatesd$b0_hat[i] = summary(model)$coefficients[1, 1]
  estimatesd$b1_hat[i] = summary(model)$coefficients[2, 1]
}

dmean_b0 = mean(estimatesd$b0_hat)
dmean_b1 = mean(estimatesd$b1_hat)

dmedian_b0 = median(estimatesd$b0_hat)
dmedian_b1 = median(estimatesd$b1_hat)
head(estimatesd)
tail(estimatesd)
# Part E
# cohort study
mad_b1 = mad(estimates$b1_hat)
```

```r
# case-control
dmad_b1 = mad(estimatesd$b1_hat)
```