The Impact of Variable Leakage Rates on OE Ratios a Logistic Model Simulation

A cohort of **four hospitals** with 1,000 randomized cases each is generated with a dichotomous outcome variable **y = 30 day readmission status** and a single dichotomous risk factor **x**. Overall, the **prevalence of y is 20%** *(readmission rate)* and the **prevalence of x is 10%**. Varying portions of each of hospital’s readmissions occurr at non-index hospitals and are referred to as **Leakage Rates = 0%, 10%, 20% and 40%.** A Spearman rank (rho) **correlation between y and x = 0.45** is used to establish the strength of association of y with x in order to achieve an over all **ROC score of ~0.66** for a logistic model: *glm(y ~ x,family = binomial(link=‘logit’))*.

- Note: the x and y prevalences (p & q), leakage rates, # hospitals, sample sizes (n.sim) and rho can all be customized.

rm(list=ls())  
library(tidyverse)

## -- Attaching packages ------------------------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.7  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.2.1 v forcats 0.3.0

## -- Conflicts --------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(kableExtra)

## Warning: package 'kableExtra' was built under R version 3.5.2

# This snippet generates the data, called once for each hospital  
# All hospitals: y prevalence (q) ~ 20%; x prevalence (p) ~10%, N=1000  
#-----------------------------------------  
p=0.1; q=0.2; rho=0.45; n=1; n.sim <- 1000  
# Compute Pr(0,0) from rho, p=Pr(X=1), and q=Pr(Y=1).  
a <- function(rho, p, q) {  
 a.0 <- rho \* sqrt(p\*q\*(1-p)\*(1-q)) + (1-p)\*(1-q)  
 prob <- c(`(0,0)`=a.0, `(1,0)`=1-q-a.0, `(0,1)`=1-p-a.0, `(1,1)`=a.0+p+q-1)  
 if (min(prob) < 0) {  
 print(prob)  
 stop("Error: a probability is negative.")  
 }  
 u <- sample.int(4, n.sim \* n, replace=TRUE, prob=prob)  
 return(u)  
}  
#-----------------------------------------

**Hospital 1:** Generate 1000 random sample x,y cases, with leakage rate = 0%

set.seed(17);u <- a(rho, p, q);y <- floor((u-1)/2);x <- 1 - u %% 2  
Hosp1 <- data.frame(hosp=1,x=x,y=y)  
Hosp1$y\_same\_Hosp <-Hosp1$y \* sample(c(rep(1,10)),n.sim,replace=T) # 0% leakage  
Hosp1$Leakage = rep(mean(Hosp1$y) - mean(Hosp1$y\_same\_Hosp),1000)

**Hospital 2:** Generate 1000 random sample x,y cases, with leakage rate = 10%

set.seed(20);u <- a(rho, p, q);y <- floor((u-1)/2);x <- 1 - u %% 2  
Hosp2 <- data.frame(hosp=2,x=x,y=y)  
Hosp2$y\_same\_Hosp <-Hosp2$y \* sample(c(rep(1,9),rep(0,1)),n.sim,replace=T) #10% leakage  
Hosp2$Leakage = rep(mean(Hosp2$y) - mean(Hosp2$y\_same\_Hosp),1000)

**Hospital 3:** Generate 1000 random sample x,y cases, with leakage rate = 20%

set.seed(21);u <- a(rho, p, q);y <- floor((u-1)/2);x <- 1 - u %% 2  
Hosp3 <- data.frame(hosp=3,x=x,y=y)  
Hosp3$y\_same\_Hosp <-Hosp3$y \* sample(c(rep(1,8),rep(0,2)),n.sim,replace=T) #20% leakage  
Hosp3$Leakage = rep(mean(Hosp3$y) - mean(Hosp3$y\_same\_Hosp),1000)

**Hospital 4:** Generate 1000 random sample x,y cases, with leakage rate = 20%

set.seed(25);u <- a(rho, p, q);y <- floor((u-1)/2);x <- 1 - u %% 2  
Hosp4 <- data.frame(hosp=4,x=x,y=y)  
Hosp4$y\_same\_Hosp <-Hosp4$y \* sample(c(rep(1,6),rep(0,4)),n.sim,replace=T) #40% leakage  
Hosp4$Leakage = rep(mean(Hosp4$y) - mean(Hosp4$y\_same\_Hosp),1000)

**Model using all data**

Hosp\_All <- rbind(Hosp1, Hosp2, Hosp3, Hosp4)  
 m1 <- glm(y ~ x , data = Hosp\_All, family = binomial(link='logit'))  
 pROC::roc(response = Hosp\_All$y, predictor = fitted(m1))

##   
## Call:  
## roc.default(response = Hosp\_All$y, predictor = fitted(m1))  
##   
## Data: fitted(m1) in 3196 controls (Hosp\_All$y 0) < 804 cases (Hosp\_All$y 1).  
## Area under the curve: 0.6731

coef(summary(m1))[,'Pr(>|z|)']

## (Intercept) x   
## 0.000000e+00 7.400952e-122

m1$coefficients

## (Intercept) x   
## -1.826177 2.755497

b0=m1$coefficients[1]  
 b1=m1$coefficients[2]

Fit for all hospitals, **evaluted using *all* captured outcomes.**

roc0 <- pROC::roc(response = Hosp\_All$y, predictor = fitted(m1))  
 roc0

##   
## Call:  
## roc.default(response = Hosp\_All$y, predictor = fitted(m1))  
##   
## Data: fitted(m1) in 3196 controls (Hosp\_All$y 0) < 804 cases (Hosp\_All$y 1).  
## Area under the curve: 0.6731

obs\_rate = sum(Hosp\_All$y)/nrow(Hosp\_All)  
 exp\_rate = sum(exp(b0+b1\*Hosp\_All$x)/(1+exp(b0+b1\*Hosp\_All$x)))/nrow(Hosp\_All)  
 oe\_ratio = obs\_rate/exp\_rate   
 oe\_ratio # should = 1.00

## [1] 1

Fit using same model but **evaluated assuming captured outcomes are same hospital only**

roc1 <- pROC::roc(response = Hosp\_All$y\_same\_Hosp, predictor = fitted(m1))  
 roc1

##   
## Call:  
## roc.default(response = Hosp\_All$y\_same\_Hosp, predictor = fitted(m1))  
##   
## Data: fitted(m1) in 3333 controls (Hosp\_All$y\_same\_Hosp 0) < 667 cases (Hosp\_All$y\_same\_Hosp 1).  
## Area under the curve: 0.6692

obs\_rate\_same = sum(Hosp\_All$y\_same\_Hosp)/nrow(Hosp\_All)  
 exp\_rate = sum(exp(b0+b1\*Hosp\_All$x)/(1+exp(b0+b1\*Hosp\_All$x)))/nrow(Hosp\_All)  
 oe\_ratio\_same = obs\_rate\_same/exp\_rate   
 oe\_ratio\_same # will be something < 1.00

## [1] 0.829602

results <- Hosp\_All %>% group\_by(hosp) %>%  
 summarize(obs\_rate=round(mean(y),2),  
 exp\_rate=round(mean(exp(b0+b1\*x)/(1+exp(b0+b1\*x))),2),  
 OE\_ratio=round(obs\_rate/exp\_rate,2),  
 obs\_rate\_same=round(mean(y\_same\_Hosp),2),  
 OE\_Ratio\_same=round(obs\_rate\_same/exp\_rate,2))  
knitr::kable(results, caption = "Table of Results",  
 col.names = c("Hosp","Obs Rate","Exp Rate","OE Ratio","Obs Rate (index only) ","OE Ratio (index only)"))

Table of Results

Hosp

Obs Rate

Exp Rate

OE Ratio

Obs Rate (index only)

OE Ratio (index only)

1

0.20

0.2

1.00

0.20

1.00

2

0.20

0.2

1.00

0.18

0.90

3

0.20

0.2

1.00

0.16

0.80

4

0.21

0.2

1.05

0.13

0.65