HW4

109354003 統研所二 吳書恆

January 12, 2022

1. Generate 500 samples of (X, Y) with $X \sim \text{Ber}(0.5)$ and $Y \sim \text{Ber}(0.3 * X + 0.6 * (1 - X))$. Test $H_0 : p_0 = p_1$, where $p_x = P(Y = 1|X = x)$.

   According to the question,

   $$Y|X = 0 \sim \text{Ber}(0.6) \text{ and } Y|X = 1 \sim \text{Ber}(0.3).$$

   Using R to generate samples and compute Chi-squared test.

```
> set.seed(1234)
> X <- rbinom(500, 1, 0.5)
> Y <- rbinom(500, 1, 0.3*X+0.6*(1-X))
> (XY.tab <- table(X, Y))
   Y
X     0   1
  0 100 143
  1 176  81
> chisq.test(XY.tab)

    Pearson's Chi-squared test with Yates' continuity correction

data:  XY.tab
X-squared = 36.629, df = 1, p-value = 1.429e-09
```

   Since $\chi^2 = 36.629$ and p-value < .001, reject $H0$. That's acceptable, because the Y distribution is different depended on X.

2. Generate W with $P(W = 1|X = 1) = 0.9$ and $P(W = 1|X = 0) = 0.75$.

   (a) Does $W|X = x$ have anything to do with Y?

   According to the question, the probability of $W|X = x$ is given and it doesn't contain any information about Y. Moreover, the information about Y can be completely determined by X. Hence, $W|X = x$ is independent to Y.

   (b) Find $(\theta_{1|1}, \theta_{1|0})$.

   Use simulation data (W, X, Y) to be true data, the true values of $(\theta_{1|1}, \theta_{1|0})$ can be calculated as follow.

```
> W <- rbinom(500, 1, 0.9*X+0.75*(1-X))
> (WXY.tab <- table(W, X, Y))
W, X, Y = 0
     0   1
  0  25  17
  1  75 159

W, X, Y = 1
     0   1
  0  29   9
  1 114  72
```

$$\theta_{1|1} = P(W=1|X=1) = (159+72)/(17+159+9+72) = 0.899$$

$$\theta_{1|0} = P(W=1|X=0) = (75+114)/(25+75+29+114) = 0.778$$

(c) Find $\frac{n_{11}}{n_{.1}}$ and compare with $\alpha_1\theta_{1|1} + (1-\alpha_1)\theta_{1|0}$.

$$\frac{n_{11}}{n_{.1}} = (114+72)/(29+9+114+72) = 0.830$$

And,

$$\alpha_1 = P(X=1|Y=1) = (9+72)/(29+9+114+72) = 0.362$$

$$\alpha_1\theta_{1|1} + (1-\alpha_1)\theta_{1|0} = 0.362 \times 0.899 + (1-0.362) \times 0.778 = 0.822$$

The values of $\frac{n_{11}}{n_{.1}}$ and $\alpha_1\theta_{1|1} + (1-\alpha_1)\theta_{1|0}$ are 0.830 and 0.822 respectively, which are close to each other.

(d) Find $\frac{n_{11}}{n_{.1}} - \frac{n_{10}}{n_{.0}}$ and compare with $(\alpha_1-\alpha_0)(\theta_{1|1}+\theta_{0|0}-1)$.

$$\frac{n_{10}}{n_{.0}} = (75+159)/(25+17+75+159) = 0.848$$

$$\frac{n_{11}}{n_{.1}} - \frac{n_{10}}{n_{.0}} = 0.830 - 0.847 = -0.017$$

And,

$$\alpha_0 = P(X=1|Y=0) = (17+159)/(25+17+75+159) = 0.638$$

$$(\alpha_1-\alpha_0)(\theta_{1|1}+\theta_{0|0}-1) = (\alpha_1-\alpha_0)(\theta_{1|1}-\theta_{1|0}) = -0.033$$

The values of $\frac{n_{11}}{n_{.1}} - \frac{n_{10}}{n_{.0}}$ and $(\alpha_1-\alpha_0)(\theta_{1|1}+\theta_{0|0}-1)$ are -0.017 and -0.033 respectively, which are close to each other either.

3. Generate 100 external validation set of (X^v, W^v) with $X^v \sim \text{Ber}(0.4)$, $P(W^v=1|X^v=1) = 0.9$ and $P(W^v = 1|X^v=0) = 0.75$.

(a) Find $(\hat{\theta}_{1|1}, \hat{\theta}_{1|0})$ and $\hat{\alpha}_1 - \hat{\alpha}_0$.

$(\hat{\theta}_{1|1}, \hat{\theta}_{1|0})$ is calculate by simulation as follow

```
> Xv <- rbinom(100, 1, 0.4)
> Wv <- rbinom(100, 1, 0.9*Xv+0.75*(1-Xv))
> (XWv.tab <- table(Wv, Xv))
   Xv
Wv   0  1
  0 16  2
  1 51 31
> theta1 <- XWv.tab[2, 2]/sum(XWv.tab[, 2])
> theta0 <- XWv.tab[2, 1]/sum(XWv.tab[, 1])
> c(theta1, theta0)
[1] 0.9393939 0.7611940
```

$$\hat{\theta}_{1|1} = 31/(2+31) = 0.939$$

$$\hat{\theta}_{1|0} = 51/(16+51) = 0.761 = 1 - \hat{\theta}_{0|0}$$

According fomula (1) below,

$$\begin{pmatrix}\hat{\alpha}_1\\ \hat{\alpha}_0\end{pmatrix} = \begin{pmatrix}\hat{\theta}_{1|1}+\hat{\theta}_{0|0}-1 & 0\\ 0 & \hat{\theta}_{1|1}+\hat{\theta}_{0|0}\end{pmatrix}^{-1}\begin{pmatrix}\frac{n_{11}}{n_{.1}}-1+\hat{\theta}_{0|0}\\ \frac{n_{10}}{n_{.0}}-1+\hat{\theta}_{0|0}\end{pmatrix} \quad (1)$$

using simulation data to get

$$\begin{pmatrix}\hat{\alpha}_1\\ \hat{\alpha}_0\end{pmatrix} = \begin{pmatrix}0.939-0.761 & 0\\ 0 & 0.939-0.761\end{pmatrix}^{-1}\begin{pmatrix}0.830-0.761\\ 0.848-0.761\end{pmatrix} = \begin{pmatrix}0.388\\ 0.486\end{pmatrix}$$

Hence, $\hat{\alpha}_1-\hat{\alpha}_0=-0.098$.

(b) Find se($\hat{\alpha}_1-\hat{\alpha}_0$) by bootstrap($B=100$).

The bootstrap($B=100$) procedure is as follow,

```
> WY <- cbind(W, Y)
> XWv <- cbind(Wv, Xv)
> alpdiff.boot <- NULL
> for(i in 1:100){
+   WY.boot <- WY[sample(500, replace = TRUE), ]
+   WY.tab.boot <- table(WY.boot[, 1], WY.boot[, 2])
+   n1.boot <- WY.tab.boot[2, 2]/sum(WY.tab.boot[, 2])
+   n0.boot <- WY.tab.boot[2, 1]/sum(WY.tab.boot[, 1])
+   XWv.boot <- XWv[sample(100, replace = TRUE), ]
+   XWv.tab.boot <- table(XWv.boot[, 1], XWv.boot[, 2])
+   theta1.boot <- XWv.tab.boot[2, 2]/sum(XWv.tab.boot[, 2])
+   theta0.boot <- XWv.tab.boot[2, 1]/sum(XWv.tab.boot[, 1])
+   theta.boot <- theta1.boot-theta0.boot
+   a.boot <- matrix(c(theta.boot, 0, 0, theta.boot), ncol=2)
+   b.boot <- c(n1.boot, n0.boot)-theta0.boot
+   c.boot <- solve(a.boot)%*%b.boot
+   alpdiff.boot <- c(alpdiff.boot, c.boot[1]-c.boot[2])
+ }
> c(mean(alpdiff.boot), sd(alpdiff.boot))
[1] -0.1039442  0.2571436
```

Hence, the mean of bootstrap of ($\hat{\alpha}_1-\hat{\alpha}_0$)$=-0.104$ and se($\hat{\alpha}_1-\hat{\alpha}_0$)$=0.257$.

4. Choose 100 internal validation set of (X,W,Y). Find estimates of $P(X=0,Y=0)$, $P(X=1,Y=0)$, $P(X=0,Y=1)$, $P(X=1,Y=1)$ and compare with true value.

The internal validation set is randomly choose by the data in (a) and (b). The tables of internal validation set (`XWYu.tab`) and whole data (`WY.tab`) is as follow.

```
> index <- sample(500, 100,replace = FALSE)
> XWYu <- cbind(X, W, Y)[index, ]
> (WY.tab <- table(W, Y))
   Y
W     0   1
  0  42  38
  1 234 186
> WY.joint <- matrix(WY.tab/sum(WY.tab))
> (XWYu.tab <- table(XWYu[, 1], XWYu[, 2], XWYu[, 3]))
X, W, Y = 0
     0  1
  0  4 16
  1  4 33

X, W, Y = 1
     0  1
  0  5 17
  1  1 20
```

According fomula (2) below,

$$\begin{pmatrix} \hat{P}(X=0,Y=0) \\ \hat{P}(X=1,Y=0) \\ \hat{P}(X=0,Y=1) \\ \hat{P}(X=1,Y=1) \end{pmatrix} = \begin{pmatrix} P(X=0|W=0,Y=0) & P(X=0|W=1,Y=0) & 0 & 0 \\ P(X=1|W=0,Y=0) & P(X=1|W=1,Y=0) & 0 & 0 \\ 0 & 0 & P(X=0|W=0,Y=1) & P(X=0|W=1,Y=1) \\ 0 & 0 & P(X=1|W=0,Y=1) & P(X=1|W=1,Y=1) \end{pmatrix} \times \begin{pmatrix} \hat{P}(W=0,Y=0) \\ \hat{P}(W=1,Y=0) \\ \hat{P}(W=0,Y=1) \\ \hat{P}(W=1,Y=1) \end{pmatrix} \tag{2}$$

using simulation data to get

$$\begin{pmatrix} \hat{P}(X=0,Y=0) \\ \hat{P}(X=1,Y=0) \\ \hat{P}(X=0,Y=1) \\ \hat{P}(X=1,Y=1) \end{pmatrix} = \begin{pmatrix} 0.5 & 0.327 & 0 & 0 \\ 0.5 & 0.673 & 0 & 0 \\ 0 & 0 & 0.833 & 0.459 \\ 0 & 0 & 0.167 & 0.541 \end{pmatrix} \begin{pmatrix} 0.084 \\ 0.468 \\ 0.076 \\ 0.372 \end{pmatrix} = \begin{pmatrix} 0.194 \\ 0.357 \\ 0.234 \\ 0.213 \end{pmatrix}$$

Compare to question 1.,

$$\begin{pmatrix} P(X=0,Y=0) \\ P(X=1,Y=0) \\ P(X=0,Y=1) \\ P(X=1,Y=1) \end{pmatrix} = \begin{pmatrix} 0.200 \\ 0.350 \\ 0.300 \\ 0.150 \end{pmatrix}$$