

CAUSAL CONTRASTS FROM OUTCOME REGRESSION

In a randomized experimental study, inference concerning the causal effect of binary variable Z on Y can be made by direct comparison of sample averages. Suppose that $Z \sim \text{Bernoulli}(p)$ for $0 < p < 1$. Two estimators of the causal contrast $\delta = \mathbb{E}^\mathcal{E}[Y(1) - Y(0)]$ can be proposed if the conditional mean model

$$\mathbb{E}_{Y|X,Z}^\mathcal{E}[Y|X = x, Z = z] = \mathbb{E}_{Y|X,Z}^\mathcal{O}[Y|X = x, Z = z] = \mu(x, z)$$

is presumed known. We have

$$\hat{\delta}_{OR} = \frac{1}{n} \sum_{i=1}^n \mu(X_i, 1) - \frac{1}{n} \sum_{i=1}^n \mu(X_i, 0) \quad (1)$$

and

$$\hat{\delta}_{OR2} = \frac{1}{n\hat{p}} \sum_{i=1}^n Z_i \mu(X_i, Z_i) - \frac{1}{n(1-\hat{p})} \sum_{i=1}^n (1 - Z_i) \mu(X_i, Z_i) \quad (2)$$

where $\hat{p} = n^{-1} \sum_{i=1}^n Z_i$. In the simulation below, we assume that $X \sim \text{Normal}(1, 1)$ and

$$Y|X = x, Z = z \sim \text{Normal}(0.5x + \delta z, 1)$$

so that $\mathbb{E}_{Y|X,Z}^\mathcal{E}[Y|X = x, Z = z] = 0.5x + \delta z$, and

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X^\mathcal{E} \left[\mathbb{E}_{Y|X,Z}^\mathcal{E}[Y|X, Z = 1] - \mathbb{E}_{Y|X,Z}^\mathcal{E}[Y|X, Z = 0] \right] = \delta.$$

In this case, $\mu(X, 1) - \mu(X, 0) = \delta$ which does not depend on X , so we can estimate δ without error using δ_{OR}

```
set.seed(23)
nreps<-1000;n<-100;p<-0.5;delta<-2
ests.mat<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
  X<-rnorm(n,1,1)
  Z<-rbinom(n,1,p)
  Y<-rnorm(n,delta*Z+0.5*X,1)
  p.hat<-mean(Z)
  ests.mat[irep,1]<-mean(delta+0.5*X)-mean(0.5*X)
  ests.mat[irep,2]<-sum(Z*(delta*Z+0.5*X))/(n*p.hat)-sum((1-Z)*(delta*Z+0.5*X))/(n*(1-p.hat))
}
apply(ests.mat,2,var)

+ [1] 7.254914e-33 9.881624e-03
```

Suppose now that

$$\mathbb{E}_{Y|X,Z}^{\mathcal{E}}[Y|X = x, Z = z] = 0.5x + \delta z + 2xz$$

so that

$$\mathbb{E}_{Y|X,Z}^{\mathcal{E}}[Y|X = x, Z = 1] - \mathbb{E}_{Y|X,Z}^{\mathcal{E}}[Y|X = x, Z = 0] = \delta + 2x.$$

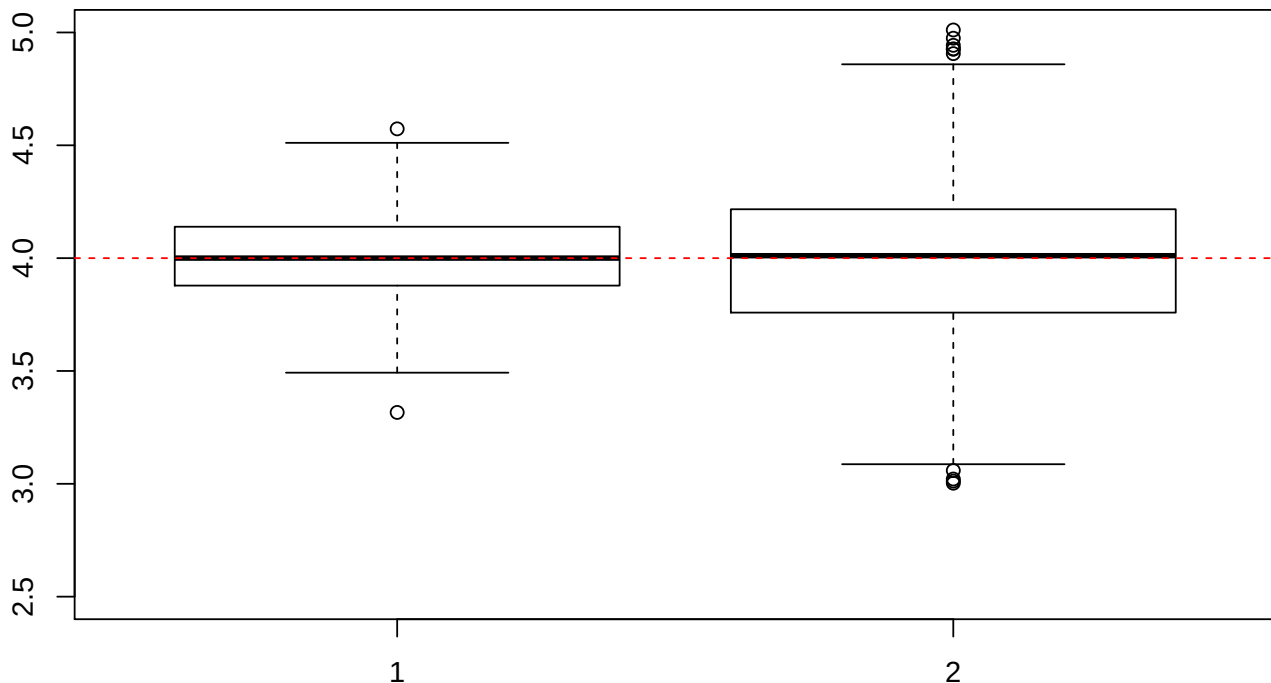
and thus

$$\mathbb{E}^{\mathcal{E}}[Y(1) - Y(0)] = \mathbb{E}^{\mathcal{E}}[\delta + 2X] = \delta + 2\mathbb{E}^{\mathcal{E}}[X] = \delta + 2$$

```
set.seed(23)
nreps<-1000;n<-100;p<-0.5;delta<-2
ests.mat<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
  X<-rnorm(n,1,1)
  Z<-rbinom(n,1,p)
  Y<-rnorm(n,delta*Z+0.5*X + 2*X*Z,1)
  p.hat<-mean(Z)
  ests.mat[irep,1]<-mean(delta+0.5*X+2*X)-mean(0.5*X)
  ests.mat[irep,2]<-sum(Z*(delta*Z+0.5*X+ 2*X*Z))/(n*p.hat)-
    sum((1-Z)*(delta*Z+0.5*X+2*X*Z))/(n*(1-p.hat))
}
apply(ests.mat,2,var)

+ [1] 0.03626926 0.12555203

par(mar=c(4,2,1,1))
boxplot(ests.mat,ylim=range(2.5,5));abline(h=delta+2,lty=2,col='red')
```



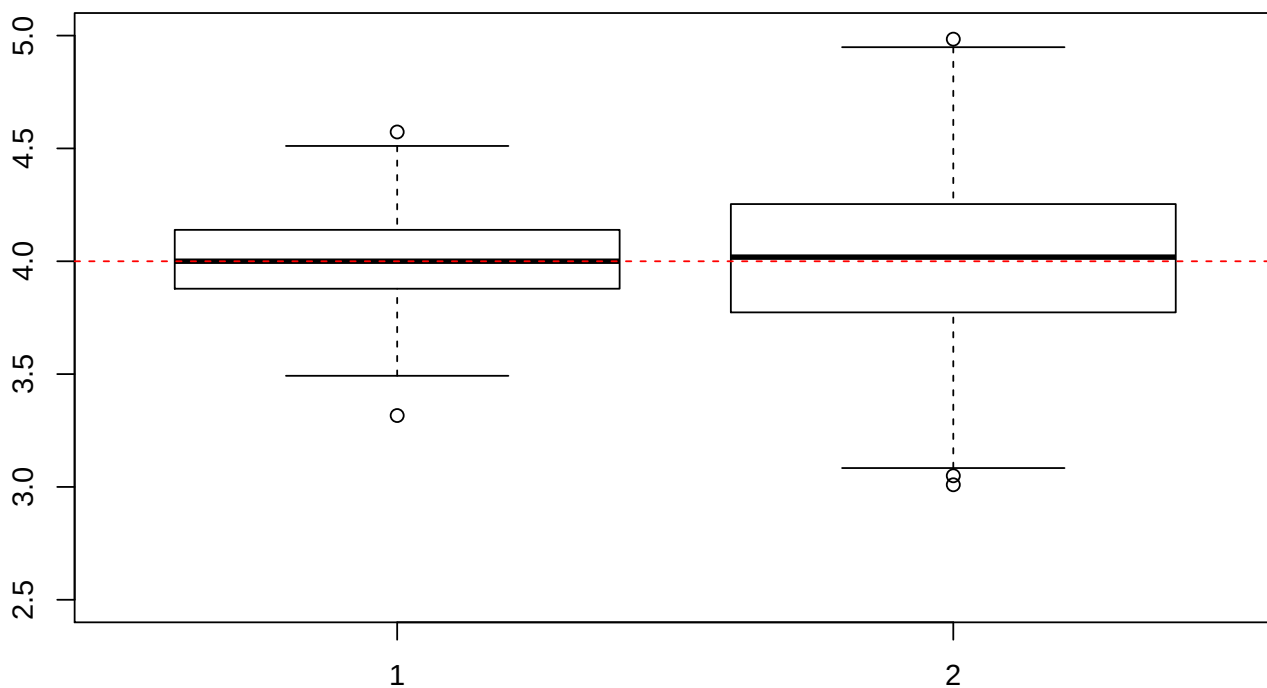
Suppose we **mis-specify** the modelled mean using

$$m(x, z) = x - 2x^2 + \delta z + 2xz$$

that is, where the dependence on z is correct, but it is not the true conditional mean. We can still unbiasedly estimate δ using the same estimator:

```
set.seed(23)
nreps<-1000;n<-100;p<-0.5;delta<-2
ests.mat<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
  X<-rnorm(n,1,1)
  Z<-rbinom(n,1,p)
  Y<-rnorm(n,delta*Z+0.5*X + 2*Z*X,1)
  p.hat<-mean(Z)
  ests.mat[irep,1]<-mean(delta+X-X^2+2*X)-mean(X-X^2)
  ests.mat[irep,2]<-sum(Z*(delta*Z+X-X^2+2*Z*X))/(n*p.hat)-
    sum((1-Z)*(delta*Z+X-X^2+2*Z*X))/(n*(1-p.hat))
}
apply(ests.mat,2,var)
+ [1] 0.03626926 0.11866337

par(mar=c(4,2,1,1))
boxplot(ests.mat,ylim=range(2.5,5));abline(h=delta+2,lty=2,col='red')
```



However, if we **mis-specify** the modelled mean using

$$m(x, z) = x - 2x^2 + \delta z + xz$$

that is, where the dependence on z is incorrect, and it is not the true conditional mean, we can no longer unbiasedly estimate δ using the same estimator:

```
set.seed(23)
nreps<-1000;n<-100;p<-0.5;delta<-2
ests.mat<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
  X<-rnorm(n,1,1)
  Z<-rbinom(n,1,p)
  Y<-rnorm(n,delta*Z+0.5*X + 2*Z*X,1)
  p.hat<-mean(Z)
  ests.mat[irep,1]<-mean(delta+X-X^2+X)-mean(X-X^2)
  ests.mat[irep,2]<-sum(Z*(delta*Z+X-X^2+Z*X))/(n*p.hat)-
    sum((1-Z)*(delta*Z+X-X^2+Z*X))/(n*(1-p.hat))
}
apply(ests.mat,2,var)

+ [1] 0.009067314 0.101345327

par(mar=c(4,2,1,1))
boxplot(ests.mat,ylim=range(2.5,5));abline(h=delta+2,lty=2,col='red')
```

