PROJECT 2: SOLUTIONS

The basic simulation will be based on:

- $X_1 \sim Normal(0.5, 0.4^2)$, $X_2 \sim Normal(-1, 1^2)$ independent.
- Conditional mean model is

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X_1 = x_1, X_2 = x_2, Z = z] = 1 + x_1 + x_2 + x_1x_2 + z(1 + x_1)$$

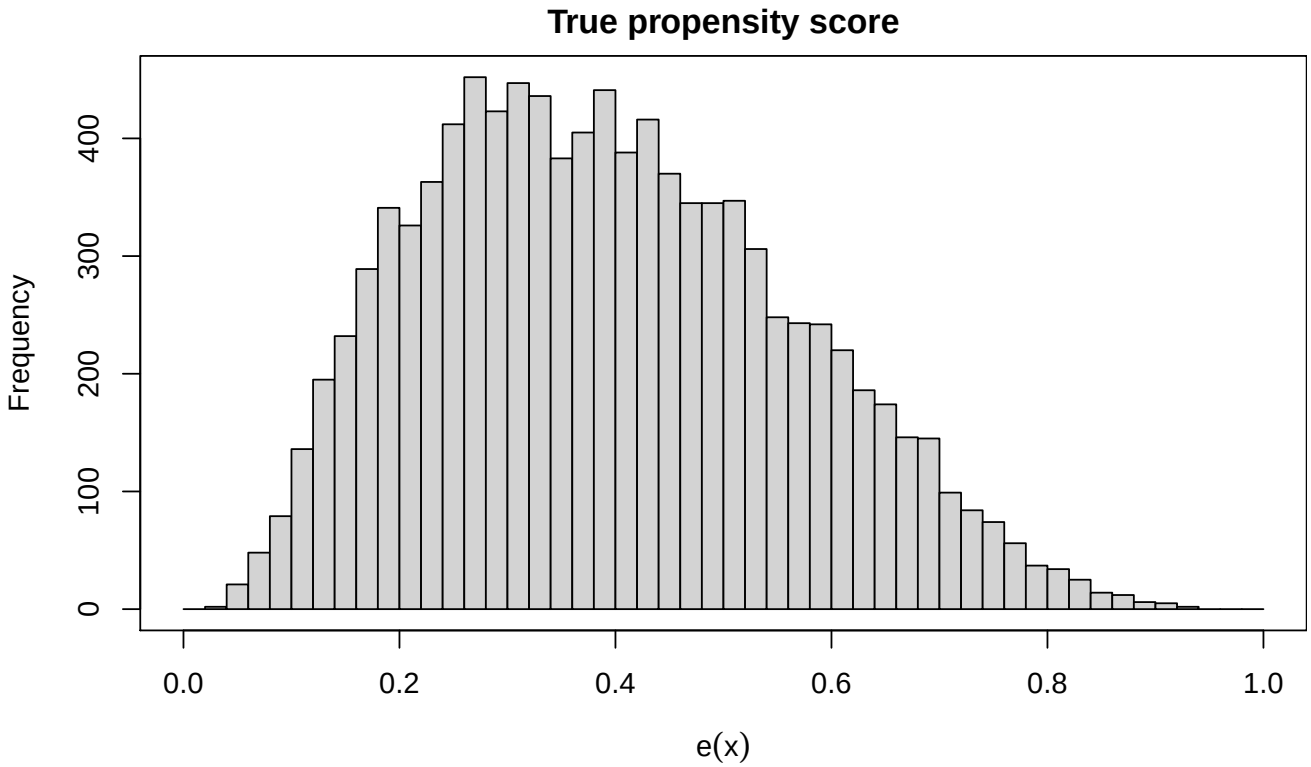
  with Normal additive errors with variance 2.5^2.
- Conditional model for $Z|X_1 = x_1, X_2 = x_2$ is $Bernoulli(e(x))$ with

$$e(x) = \frac{\exp\{-1.5 + 2x_1\}}{1 + \exp\{-1.5 + 2x_1\}}.$$

That is, X_1 is a confounder, but X_2 is not. Here the ATE is

$$\delta = \mu(1) - \mu(0) = 1 + \mathbb{E}[X_1] = 1.5$$

```
#Calculation for large N
library(mvnfast)
set.seed(22087)
N<-10000
mu1<-0.5;sig1<-0.4
X1<-rnorm(N,mu1,sig1)
mu2<--1;sig2<-1
X2<-rnorm(N,mu2,sig2)
al<-c(-1.5,2)
expit<-function(x){return(1/(1+exp(-x)))}
Xa<-cbind(1,X1)
eX0<-expit(Xa %*% al)
Z<-rbinom(N,1,eX0)
Xb<-cbind(1,X1,X2,X1*X2,Z,Z*X1)
be<-c(1,1,1,1,1,1)
sigY<-2.5
Y<-Xb %*% be + rnorm(N)*sigY
par(mar=c(4,4,2,0))
hist(eX0,breaks=seq(0,1,by=0.02),main='True propensity score',xlab=expression(e(x)))
box()
```

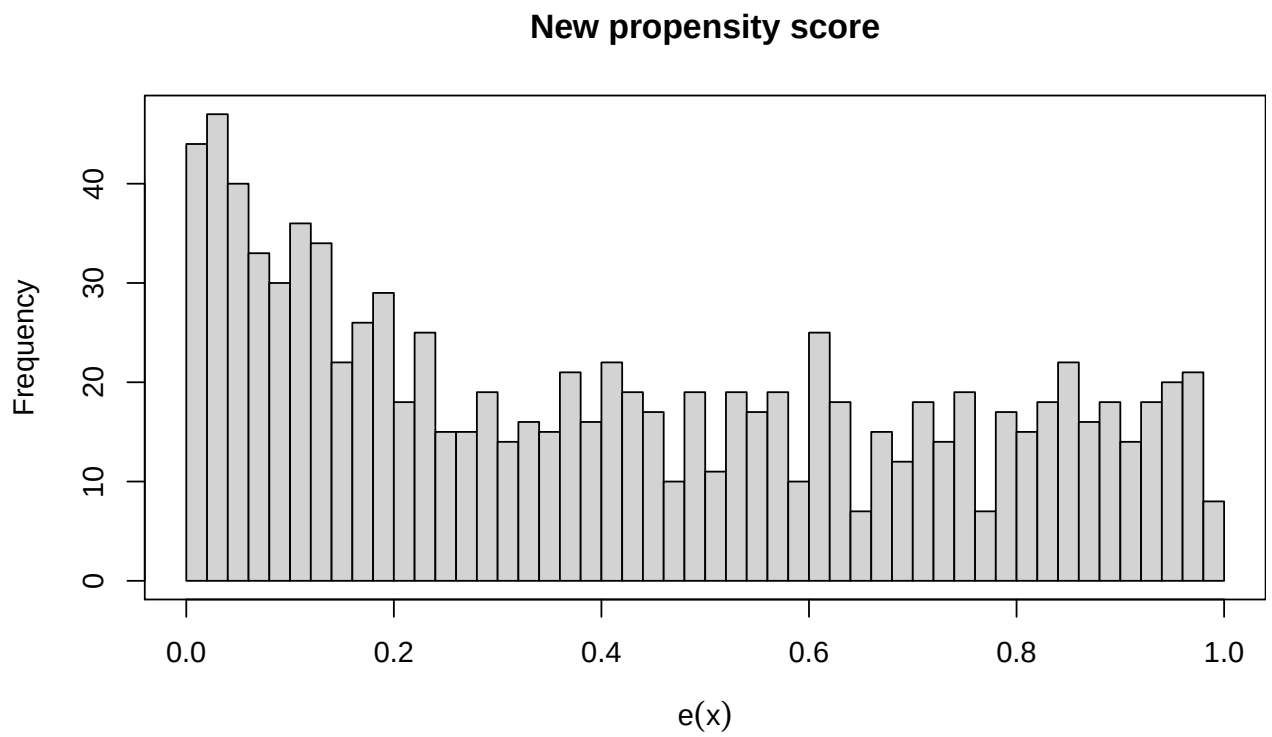



```
#Test Correct model
fit0<-lm(Y~X1+X2+X1:X2+Z+Z:X1)
mu0.0<-mean(predict(fit0,newdata=data.frame(X1=X1,X2=X2,Z=0)))
mu1.0<-mean(predict(fit0,newdata=data.frame(X1=X1,X2=X2,Z=1)))
mu1.0-mu0.0

+ [1] 1.606443
```

```
#Base Monte Carlo
nreps<-10000
n<-1000
ests0<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
    X1<-rnorm(n,mu1,sig1)
    X2<-rnorm(n,mu2,sig2)
    Xa<-cbind(1,X1)
    eX0<-expit(Xa %*% al)
    Z<-rbinom(n,1,eX0)
    Xb<-cbind(1,X1,X2,X1*X2,Z,Z*X1)
    Y<-Xb %*% be + rnorm(n)*sigY

    w0<-(1-Z)/(1-eX0)
    w1<-Z/eX0
    ests0[irep,1]<-mean(w1*Y)-mean(w0*Y)
    ests0[irep,2]<-sum(w1*Y)/sum(w1)-sum(w0*Y)/sum(w0)
}
apply(ests0,2,mean)

+ [1] 1.497866 1.499295

apply(ests0,2,var)

+ [1] 0.04533900 0.04267598
```

(a) The distribution of the propensity score influences the *variance* of the ATE estimator.

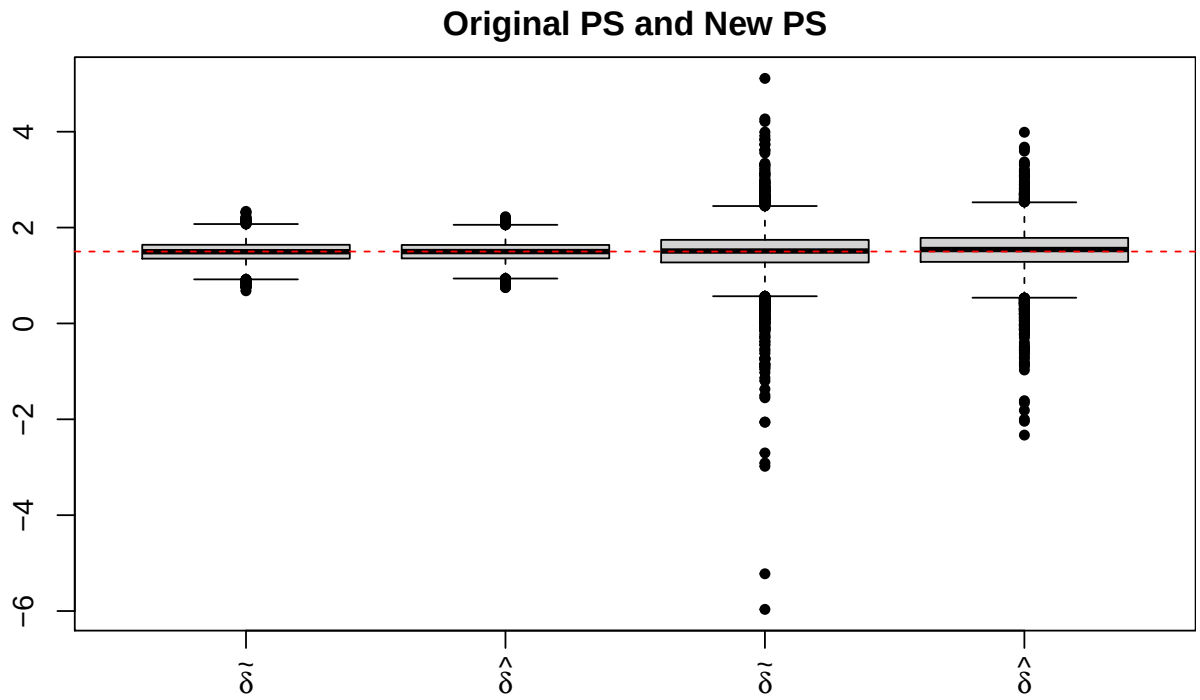
```
#Dependence on PS model
nreps<-10000
n<-1000
ests1<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
    X1<-rnorm(n,mu1,1)      #Change sigma1 to 1
    X2<-rnorm(n,mu2,sig2)
    Xa<-cbind(1,X1)
    eX0<-expit(Xa %*% al)
    Z<-rbinom(n,1,eX0)
    Xb<-cbind(1,X1,X2,X1*X2,Z,Z*X1)
    Y<-Xb %*% be + rnorm(n)*sigY

    w0<-(1-Z)/(1-eX0)
    w1<-Z/eX0
    ests1[irep,1]<-mean(w1*Y)-mean(w0*Y)
    ests1[irep,2]<-sum(w1*Y)/sum(w1)-sum(w0*Y)/sum(w0)
}
par(mar=c(4,4,4,0))
hist(eX0,breaks=seq(0,1,by=0.02),main='New propensity score',xlab=expression(e(x)))
box()
```




```
apply(ests1,2,mean)

+ [1] 1.497411 1.517667

apply(ests1,2,var)

+ [1] 0.1910117 0.1785886

apply(ests1,2,var)/apply(ests0,2,var)  #Ratio of variances compared to base case.

+ [1] 4.212967 4.184756
```

(b) Estimators based on $\widetilde{\mu}(z)$ have different *variances* to estimators based on $\widehat{\mu}(z)$.

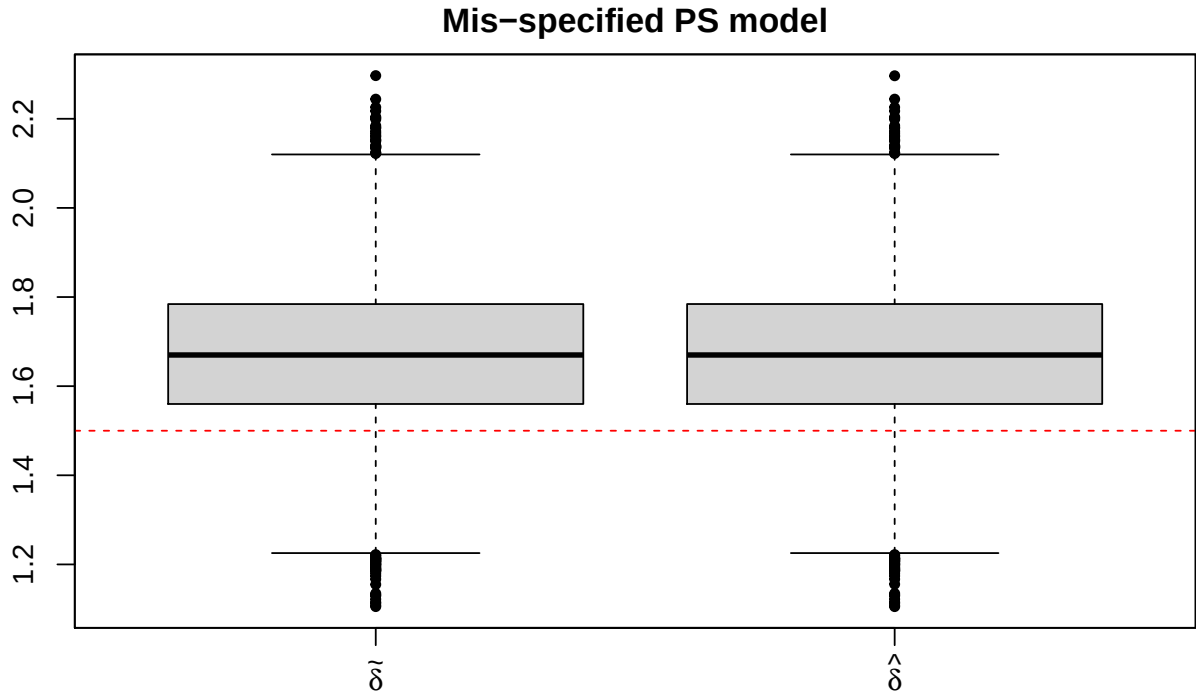
```
#From the previous results
par(mar=c(4,4,2,0))
boxplot(cbind(ests0,ests1),pch=19,cex=0.7,
    names=c(expression(tilde(delta)),expression(hat(delta)),
    expression(tilde(delta)),expression(hat(delta))))
abline(h=1.5,lty=2,col='red')
title('Original PS and New PS')
```




```
v0<-apply(ests0,2,var)
v1<-apply(ests1,2,var)
v1[1]/v0[1]

+ [1] 4.212967

v1[2]/v0[2]

+ [1] 4.184756
```

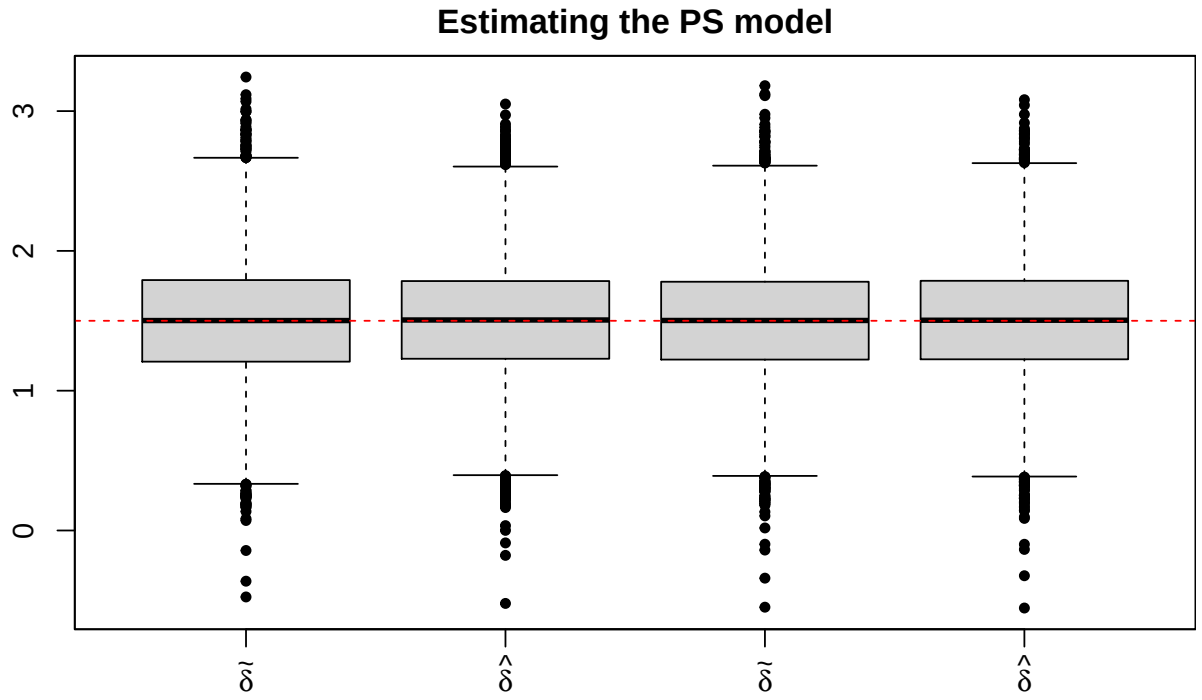
(c) Mis-specification of a propensity score model can lead to *biased (and inconsistent)* estimation of the ATE.

```
#Mis-specification
nreps<-10000
n<-1000
ests2<-matrix(0,nrow=nreps,ncol=2)
for(irep in 1:nreps){
    X1<-rnorm(n,mu1,sig1)
    X2<-rnorm(n,mu2,sig2)
    Xa<-cbind(1,X1)
    eX0<-expit(Xa %*% al)
    Z<-rbinom(n,1,eX0)
    Xb<-cbind(1,X1,X2,X1*X2,Z,Z*X1)
    Y<-Xb %*% be + rnorm(n)*sigY

    Xax<-cbind(1,X2)                        #Use X2 instead of X1
    eXx<-fitted(glm(Z~X2,family=binomial))

    w0<-(1-Z)/(1-eXx)
    w1<-Z/eXx
    ests2[irep,1]<-mean(w1*Y)-mean(w0*Y)
    ests2[irep,2]<-sum(w1*Y)/sum(w1)-sum(w0*Y)/sum(w0)
}
apply(ests2,2,mean)

+ [1] 1.670869 1.670873

par(mar=c(4,4,2,0))
boxplot(ests2,names=c(expression(tilde(delta)),expression(hat(delta))),pch=19,cex=0.7)
title('Mis-specified PS model')
abline(h=1.5,lty=2,col='red')
box()
```



(d) Estimation and use of a *parametric* propensity score model can improve the variance of an ATE estimator compared with using the *known* propensity score form.

```
#Estimating the PS
nreps<-10000
n<-250
ests3<-matrix(0,nrow=nreps,ncol=4)
for(irep in 1:nreps){
    X1<-rnorm(n,mu1,sig1)
    X2<-rnorm(n,mu2,sig2)
    Xa<-cbind(1,X1)
    eX0<-expit(Xa %*% al)
    Z<-rbinom(n,1,eX0)
    Xb<-cbind(1,X1,X2,X1*X2,Z,Z*X1)
    Y<-Xb %*% be + rnorm(n)*sigY

    w0<-(1-Z)/(1-eX0)
    w1<-Z/eX0
    ests3[irep,1]<-mean(w1*Y)-mean(w0*Y)
    ests3[irep,2]<-sum(w1*Y)/sum(w1)-sum(w0*Y)/sum(w0)

    eX<-fitted(glm(Z~X1,family=binomial))

    w0<-(1-Z)/(1-eX)
    w1<-Z/eX
    ests3[irep,3]<-mean(w1*Y)-mean(w0*Y)
    ests3[irep,4]<-sum(w1*Y)/sum(w1)-sum(w0*Y)/sum(w0)

}
apply(ests3,2,var)

+ [1] 0.1863788 0.1734538 0.1762408 0.1756372

par(mar=c(4,4,2,0))
boxplot(ests3,names=c(expression(tilde(delta)),expression(hat(delta)),
    expression(tilde(delta)),expression(hat(delta))),pch=19,cex=0.7)
title('Estimating the PS model')
abline(h=1.5,lty=2,col='red')
box()
```



(e) Propensity score models need only be constructed from *confounders* rather than all predictors. We now change the data generating model

- $X_1 \sim Normal(0.5, 0.4^2)$, $X_2 \sim Normal(-1, 1^2)$ independent.
- Conditional mean model is

$$\mathbb{E}^{\circ}_{Y|X,Z}[Y|X_1 = x_1, X_2 = x_2, Z = z] = 1 + x_1 + z(1 + x_1)$$

  with Normal additive errors with variance 2.5^2.
- Conditional model for $Z|X_1 = x_1, X_2 = x_2$ is $Bernoulli(e(x))$ with

$$e(x) = \frac{\exp\{-1.5 + x_1 + x_2\}}{1 + \exp\{-1.5 + x_1 + x_2\}}.$$

```
#Estimating the PS
nreps<-10000
n<-1000
al<-c(-1.5,1,1)
ests4<-matrix(0,nrow=nreps,ncol=4)
for(irep in 1:nreps){
    X1<-rnorm(n,mu1,sig1)
    X2<-rnorm(n,mu2,sig2)
    Xa<-cbind(1,X1,X2)
    eX0<-expit(Xa %*% al)
    Z<-rbinom(n,1,eX0)
    Xb<-cbind(1,X1,Z,Z*X1)
    Y<-Xb %*% be[1:4] + rnorm(n)*sigY

    eX1<-fitted(glm(Z~X1+X2,family=binomial))

    w0<-(1-Z)/(1-eX1)
    w1<-Z/eX1
```

```
    ests4[irep,1]<-mean(w1*Y)-mean(w0*Y)
    ests4[irep,2]<-sum(w1*Y)/sum(w1)-sum(w0*Y)/sum(w0)

    eX<-fitted(glm(Z~X1,family=binomial))

    w0<-(1-Z)/(1-eX)
    w1<-Z/eX
    ests4[irep,3]<-mean(w1*Y)-mean(w0*Y)
    ests4[irep,4]<-sum(w1*Y)/sum(w1)-sum(w0*Y)/sum(w0)

}
apply(ests4,2,mean)

+ [1] 1.495577 1.500804 1.499152 1.500930

apply(ests4,2,var)

+ [1] 0.12012116 0.10085011 0.04966831 0.05034470

par(mar=c(4,4,2,0))
boxplot(ests4,names=c(expression(tilde(delta)[12]),expression(hat(delta)[12]),
    expression(tilde(delta)[1]),expression(hat(delta)[1])),pch=19,cex=0.7)
title('Estimating the PS model only using confounder X1')
abline(h=1.5,lty=2,col='red')
box()
```