MAS6012-370- Exercise 1

Andrew Saul - 170138286

26-Feb-19

1 Question 1

Let $Y = X_1^2 + 2X_2X_3$, where

$X_1 \sim N(6, 9)$, $X_2 \sim U[9, 14]$, $X_3 \sim N(9.1)$, and X_1, X_2 and X_3 are independent.

As X_2 has uniform distribution, $\mathbb{E}X_2 = \frac{14+9}{2} = 11.5$ and $Var(X_2) = \frac{(14-9)^2}{12} = \frac{25}{12}$.
The task is to find which of the conditional variances $Var[Y|X_i]$, where $i = 1, 2, 3$, is largest. Knowing the true quantity of the variable X_i that results in largest variance $Var[Y|X_i]$, will result in the smallest variance of Y.

1.1 Var[E(Y|X__i)]

To find $Var[E(Y|X_i)]$ we need first to find $\mathbb{E}(X|X_i)$.

1.1.1 Var[E(Y|X__1)]

Keeping X_1 fixed, and finding the expectation for X_2 and X_3 results in

$$\begin{aligned}\mathbb{E}[Y|X_1] &= X_1^2 + 2\mathbb{E}(X_2)\mathbb{E}(X_3) \\ &= X_1^2 + 2(11.5)(9) \\ &= X_1^2 + 207\end{aligned}$$

Now $Var(\mathbb{E}[Y|X_1]) = Var(X_1^2 + 207) = Var(X_1^2)$.

As $Var(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$, then

$$\begin{aligned}Var(X^2) &= \mathbb{E}(X^2)^2 - [\mathbb{E}(X)^2]^2 \\ &= \mathbb{E}(X^4) - [\mathbb{E}(X)^2]^2\end{aligned}$$

The non-central 4th moment is given in the computer experiment notes (p.10) as

$$\mathbb{E}(X^4) = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$$

The non-central 2nd moment can be derived from the variance of X.

$$\begin{aligned}Var(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \\ \implies \mathbb{E}(X^2) &= Var(X) + [\mathbb{E}(X)]^2 \\ &= \mu^2 + \sigma^2\end{aligned}$$

Therefore

$$\begin{aligned} Var(X^2) &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 - [\mu^2 + \sigma^2]^2 \\ &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 - [\mu^4 - 2\mu^2\sigma^2 + \sigma^4] \\ &= 8\mu^2\sigma^2 - 2(\sigma^2)^2 \end{aligned}$$

We know that for X_1, $\mu = 6$ and $\sigma^2 = 9$. As $Var(X^2) = Var(\mathbb{E}[Y|X_1])$, we have

$$\begin{aligned} Var(\mathbb{E}[Y|X_1]) &= 8(6)^2(9) - 2(9)^2 \\ &= 2592 - 162 \\ &= 2430 \end{aligned}$$

So $Var(\mathbb{E}[Y|X_1]) = 2430$

1.1.2 Var[E(Y|X_2)]

Keeping X_2 fixed, and finding the expectation for X_1 and X_3 results in

$$\mathbb{E}[Y|X_2] = \mathbb{E}(X_1^2) + 2X_2\mathbb{E}(X_3)$$

We know that $Var(X_1) = \mathbb{E}(X_1^2) - [\mathbb{E}(X_1)]^2$. Therefore $\mathbb{E}(X_1^2) = Var(X_1) + [\mathbb{E}(X_1)]^2$. Also, we know that $Var(X_1) = 9$, $\mathbb{E}(X_1) = 6$ and $\mathbb{E}(X_3) = 9$. So we now have

$$\begin{aligned} \mathbb{E}[Y|X_2] &= \mathbb{E}(X_1^2) + 2X_2\mathbb{E}(X_3) \\ &= Var(X_1) + [\mathbb{E}(X_1)]^2 + 2X_2\mathbb{E}(X_3) \\ &= 9 + (6)^2 + 2X_2(9) \\ &= 45 + 18(X_2) \end{aligned}$$

Given that $Var(X_2) = \frac{25}{12}$ we have

$$\begin{aligned} Var(\mathbb{E}[Y|X_2]) &= Var(45 + 18(X_2)) \\ &= Var(18X_2) \\ &= 324 \times Var(X_2) \\ &= 324(\frac{25}{12}) \\ &= 675 \end{aligned}$$

So $Var(\mathbb{E}[Y|X_2]) = 675$

1.1.3 Var[E(Y|X_3)]

Keeping X_3 fixed, and finding the expectation for X_1 and X_2 results in

$$\mathbb{E}[Y|X_3] = \mathbb{E}(X_1^2) + 2\mathbb{E}(X_2)X_3$$

We have calculated $\mathbb{E}(X_1^2) = 45$ and $\mathbb{E}(X_2) = 11.5$. Therefore

$$\begin{aligned} \mathbb{E}[Y|X_3] &= \mathbb{E}(X_1^2) + 2\mathbb{E}(X_2)X_3 \\ &= 45 + 2(11.5)X_3 \\ &= 45 + 23X_3 \end{aligned}$$

Now, since $Var(X_3) = 1$, we have

$$\begin{aligned}
Var(\mathbb{E}[Y|X_3]) &= Var(45 + 23X_3) \\
&= Var(23X_3) \\
&= 529 \times Var(X_3) \qquad = 529(1) \\
&= 529
\end{aligned}$$

So $Var(\mathbb{E}[Y|X_3]) = 529$

1.2 Conclusion

As the greatest variation in Y occurred when X_1 was fixed (2430), as compared to fixing X_2 (675) or X_3 (529), I would suggest the best input to learn would be X_1. By learning X_1, the total variation in Y will have been reduced as much as possible.

2 Question 2

2.1 Q2a

Modified code of cantilever beam having removed input R and output S.

```
canti <- function(xx, w=4, t=2)
{

  ##########################################################################
  #
  # OUTPUTS AND INPUTS:
  #
  #  D = displacement
  # xx = c(E, X, Y)
  # w  = width (optional)
  # t  = thickness (optional)
  #
  #########################################################################


  E <- xx[1]
  X <- xx[2]
  Y <- xx[3]

  L   <- 100


  Dfact1 <- 4*(L^3) / (E*w*t)
  Dfact2 <- sqrt((Y/(t^2))^2 + (X/(w^2))^2)

  D <- Dfact1 * Dfact2

  return(D)
```

```
}
```

2.2 Q2b

Two approaches were used with MC simulation to estimate variance of displacement $D(\mathbf{x})$, where $\mathbf{x} = (\mathbf{E}, \mathbf{X}, \mathbf{Y})$. The first utilised a for loop with 100,000 draws, which permitted the cumulative variance of D(x) to be plotted. From figure 1 it can be seen that equilibrium of the MC chain is reached at around 30,000 draws. As such, $Var(D(x))$ was measured from draws 40,000 to 100,000.

```
#set number of draws for MC simulation and create empty vector to accommodate response
#variable values.  Also, as an addition, create empty set for the calculated cumulative
#variability
N1 <- 100000
Dloop <- vector("numeric",N1)
e <- vector("numeric",N1)
x <- vector("numeric",N1)
y <- vector("numeric",N1)
cum.var <- vector("numeric",N1)

#perform MC approach to estimate Var(D(x) and the cumulative variability of D(x))
set.seed(123)
cum.var[1] <- NA
for (i in 1:N1){
  e <-rnorm(1,1.9e7,1.7e6)
  x <- rnorm(1,500,100)
  y <- rnorm(1,1100,100)
Dloop[i] <- canti(c(e,x,y))
cum.var[i] <- var(Dloop[1:i])
}

#plot cumulative variability D(x)
plot( c(2:N1),cum.var[2:N1], xlab = "draw number", ylab="Cumulative variance D(x)"
      , ylim = c(0.7,1))
```

```
c("The variation in D(x) is" ,format(var(Dloop[40000:N1]), digits=3))
```

```
## [1] "The variation in D(x) is" "0.901"
```

A second method using the "apply" function and 500,000 draws were utilised to determine $Var(D(x))$.

```
#No. draws of MC simulation
N <-  500000
#define vector length for e,x,y input variables and D output variable
E <- vector("numeric",N)
X <- vector("numeric",N)
Y <- vector("numeric",N)
D <- vector("numeric",N)

#obtain values for e,x,y for each simulation draw
```

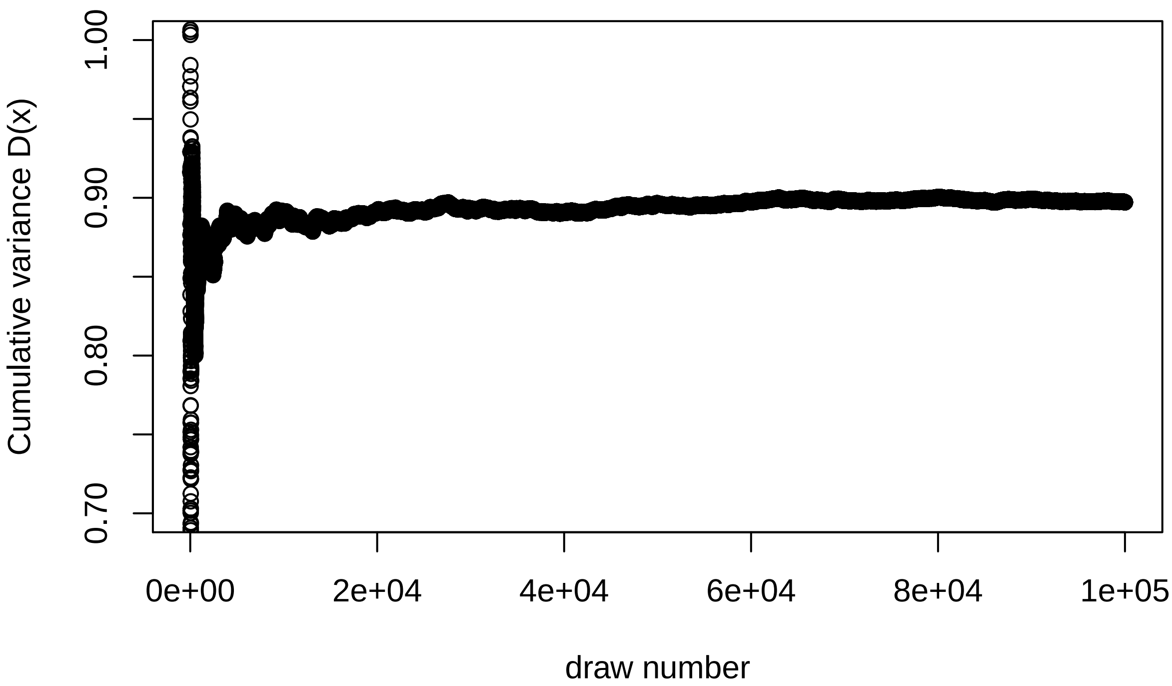

Figure 1: Plot of cumulative variance D(x) against MC draw. Equibilibrium is reached at around 30,000 draws.

```r
set.seed(123)
E <-rnorm(N,1.9e7,1.7e6)
X <- rnorm(N,500,100)
Y <- rnorm(N,1100,100)

#combine input vectors to form a matrix
m <- matrix(cbind(E,X,Y), ncol=3)

#Obtain output vector D using the canti function
DV <- apply(m,1,canti)

#Var(D)
paste("Variation of displacement from draw 40,000 to",N,"="
      ,format(var(DV[40000:N]), digits=3))
```

```
## [1] "Variation of displacement from draw 40,000 to 5e+05 = 0.892"
```

```r
paste("Variation of displacement of ",N," draws = ",format(var(DV), digits=3))
```

```
## [1] "Variation of displacement of  5e+05  draws =  0.892"
```

As 1,000,000 draws were utilised in the MC simulation, the effect of disequilibrium for the initial burn-in period had neglible effect on the overall variance of displacement $D(\mathbf{x})$. Using both for loops and vectorisation, the variance of $D(\mathbf{x})$ was between 0.89 and 0.90.

2.3 Q2c

GAM regression, using the "mgcv" library was utilised to estimate the variance of the main effects E,X,Y, and all associated interaction terms. Adding all variances of main effects and interaction terms yielded the total variance D. Taking the ratio of main effect from the total variance (D) produced the main effect index. A total of 10,000 draws was utilised for the GAM regression. Using greater than 10,000 draws had little impact on the main effects indices.

```r
#load mgcv library for GAM regression
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-24. For overview type 'help("mgcv-package")'.
```

```r
#specify variables and parameters of model
N <- 10000
E <- vector("numeric",N)
X <- vector("numeric",N)
Y <- vector("numeric",N)
D <- vector("numeric",N)

set.seed(123)
E <-rnorm(N,1.9e7,1.7e6)
X <- rnorm(N,500,100)
Y <- rnorm(N,1100,100)
```

```r
L <- 100
w=4; t=2

#equation of model
D <- (4*(L^3) / (E*w*t))*sqrt((Y/(t^2))^2 + (X/(w^2))^2)

# Estimating the variance decomposition for main effects
#and interactions

gam.model1 <- gam(D ~ s(E))
gam.model2 <- gam(D ~ s(X))
gam.model3 <- gam(D ~ s(Y))
gam.model12 <- gam(D ~ te(E,X))
gam.model13 <- gam(D ~ te(E,Y))
gam.model23 <- gam(D ~ te(X,Y))
gam.model123 <- gam(D ~ te(E,X,Y))

v.E<-var(gam.model1$fitted)
v.X<-var(gam.model2$fitted)
v.Y<-var(gam.model3$fitted)
v.EX<-var(gam.model12$fitted)
v.EY<-var(gam.model13$fitted)
v.XY<-var(gam.model23$fitted)
v.EXY<-var(gam.model123$fitted)

# Variance of components
#matrix(c(v.E, v.X, v.Y, v.EX - v.E - v.X, v.EY - v.E - v.Y,
        # v.XY - v.X - v.Y, v.EXY + v.E + v.X + v.Y - v.EX - v.EY -v.XY, var(Y) ),8,1)

#Total variance D
var.D <- v.E+v.X+v.Y+v.EX+v.EY+v.XY+v.EXY

# main effects indices
var.E.prop <- v.E/var.D
var.X.prop <- v.X/var.D
var.Y.prop <- v.Y/var.D

paste("The main effect (E) index = ", format((var.E.prop), digits=3))
```

```
## [1] "The main effect (E) index =  0.126"
```

```r
paste("The main effect (X) index = ", format((var.X.prop), digits=3))
```

```
## [1] "The main effect (X) index =  5.84e-05"
```

```r
paste("The main effect (Y) index = ", format((var.Y.prop), digits=3))
```

```
## [1] "The main effect (Y) index =  0.121"
```

The results of GAM regression indicate that the main effect (E) alone accounts for 12.6% of model variation, the main effect (Y) accounts for 12.1% of model variation, and the main effect (X) has neglible impact on model variation.