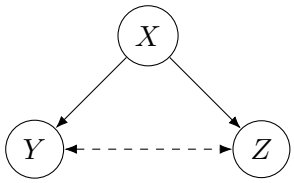
USING CAUSAL GRAPHS IN EPIDEMIOLOGICAL RESEARCH
SIMULATIONS AND EXAMPLES

1. CONDITIONAL INDEPENDENCE AND DEPENDENCE

In this example, we aim to illustrate the difference between conditional independence and dependence. Suppose we have the simple graph



where the 'bidirected' edge between Y and Z indicates a general dependence between those variables. We can read off the joint distribution

$$f_X(x) f_{Y,Z|X}(y, z|x).$$

We can generate data as follows: first, under the assumption that the variables are jointly Normally distributed, with

$$X \sim Normal(10, 5^2)$$

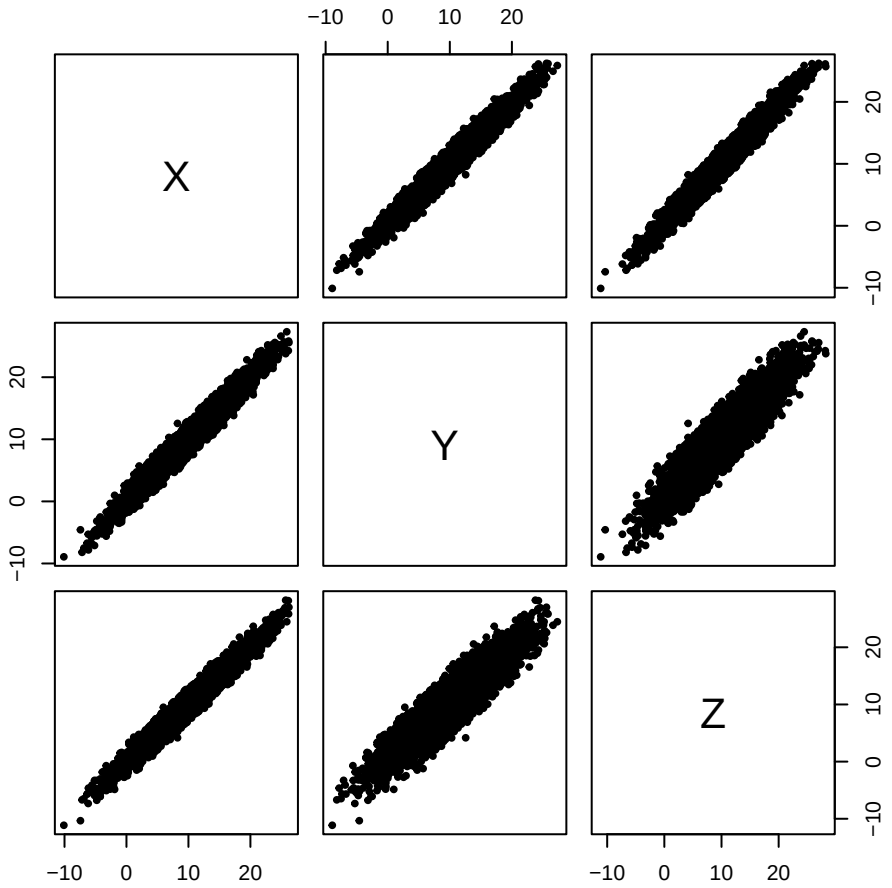
$$(Y, Z)|X = x \sim Normal\left(\begin{pmatrix} x \\ x \end{pmatrix}, \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}\right)$$

so given $X = x$, Y and Z are *dependent*, each with mean x but with correlation -0.9.

```
library(MASS)
set.seed(2111)                          #Set the random number generator seed value
n<-10000                                #Set the sample size
X<-rnorm(n,10,5)                        #Generate the X random variables
Sig.YZ<-matrix(c(1,-0.9,-0.9,1),2,2)    #Conditional variance-covariance for Y,Z given X
YZ<-mvrnorm(n,mu=c(0,0),Sigma=Sig.YZ)   #Generate the Y,Z variables
Y<-X+YZ[,1];Z<-X+YZ[,2]                 #Change the mean according to X
cor(cbind(X,Y,Z))                       #Compute the unconditional correlation

+           X         Y         Z
+ X 1.0000000 0.9805354 0.9807260
+ Y 0.9805354 1.0000000 0.9270599
+ Z 0.9807260 0.9270599 1.0000000
```

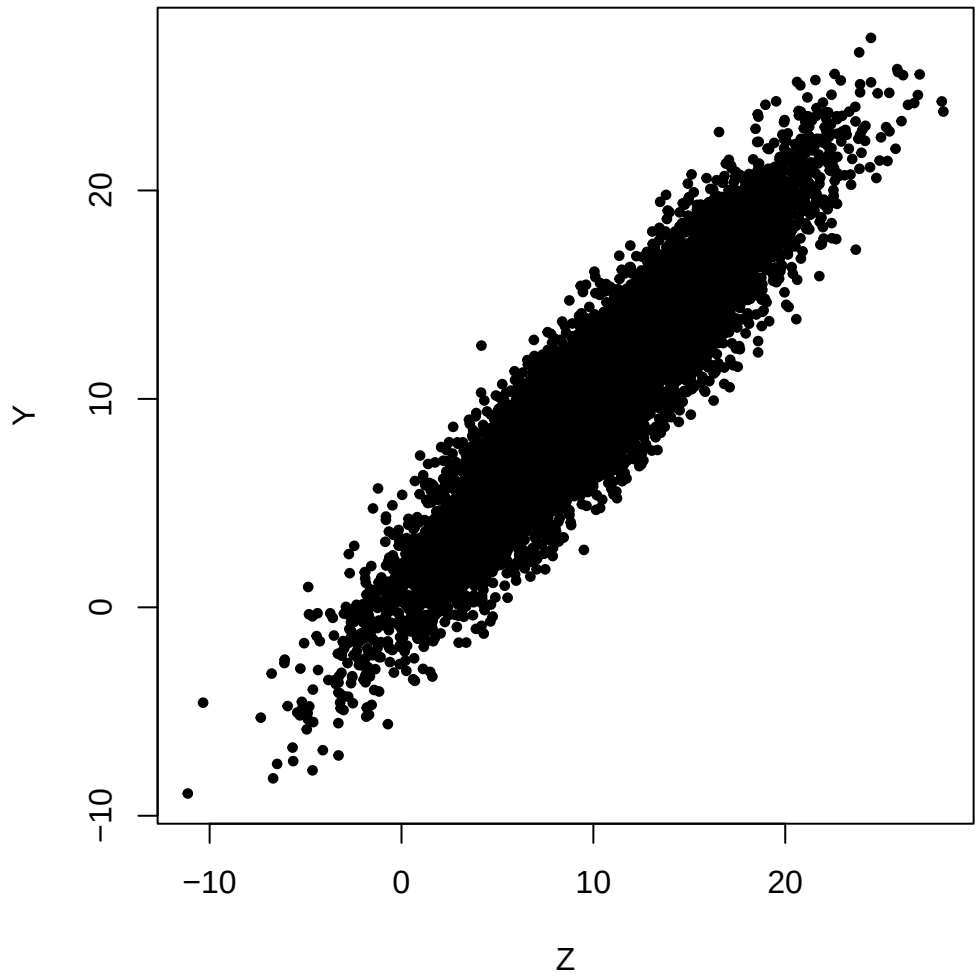
We see that marginally all the variables are positively correlated. We can see this further in a points plot

```
par(mar=c(4,4,1,0),pty='s')             #Set up the plotting margins
pairs(cbind(X,Y,Z),pch=19,cex=0.6)      #Scatterplot matrix
```



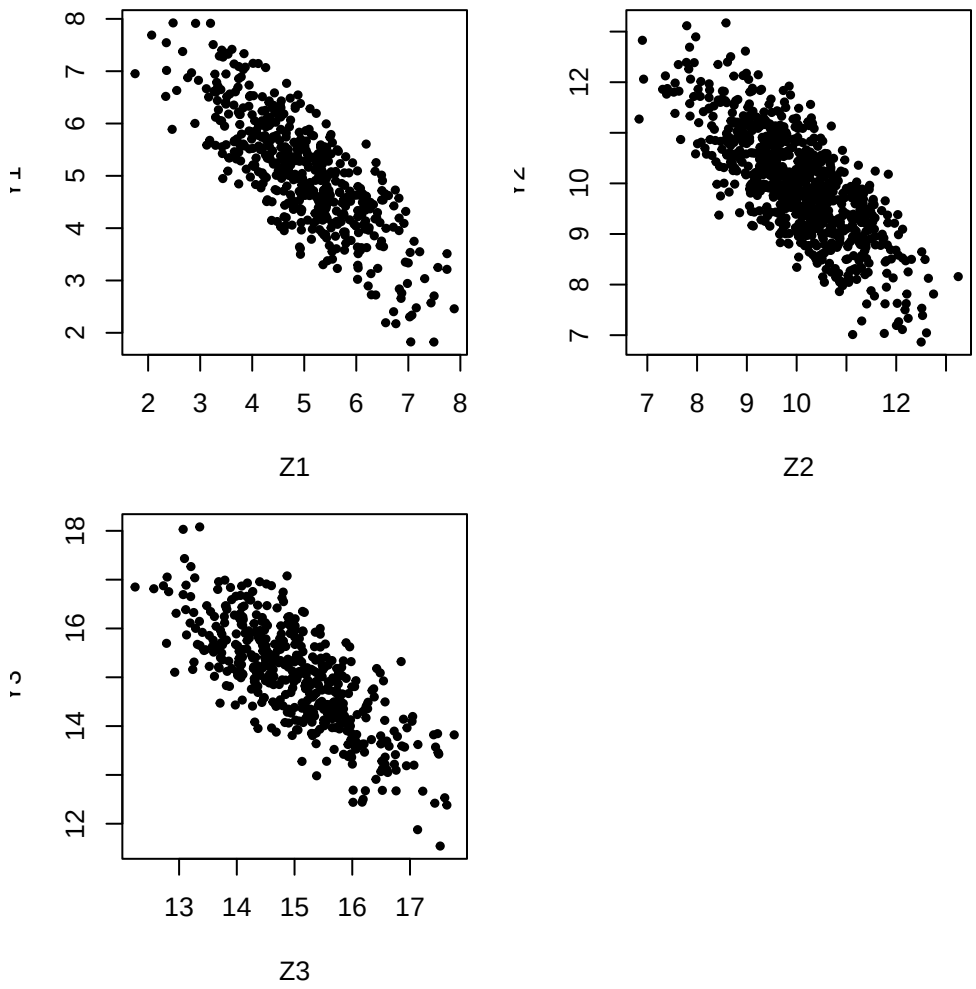
There is evidently strong positive correlation between each pair of variables, including Y and Z; this may be a surprise as conditionally, they are negatively correlated.

```
par(mar=c(4,4,1,0),pty='s')              #Set up the plotting margins
plot(Z,Y,pch=19,cex=0.6)
```



We now examine three subsets defined by different ranges of X:

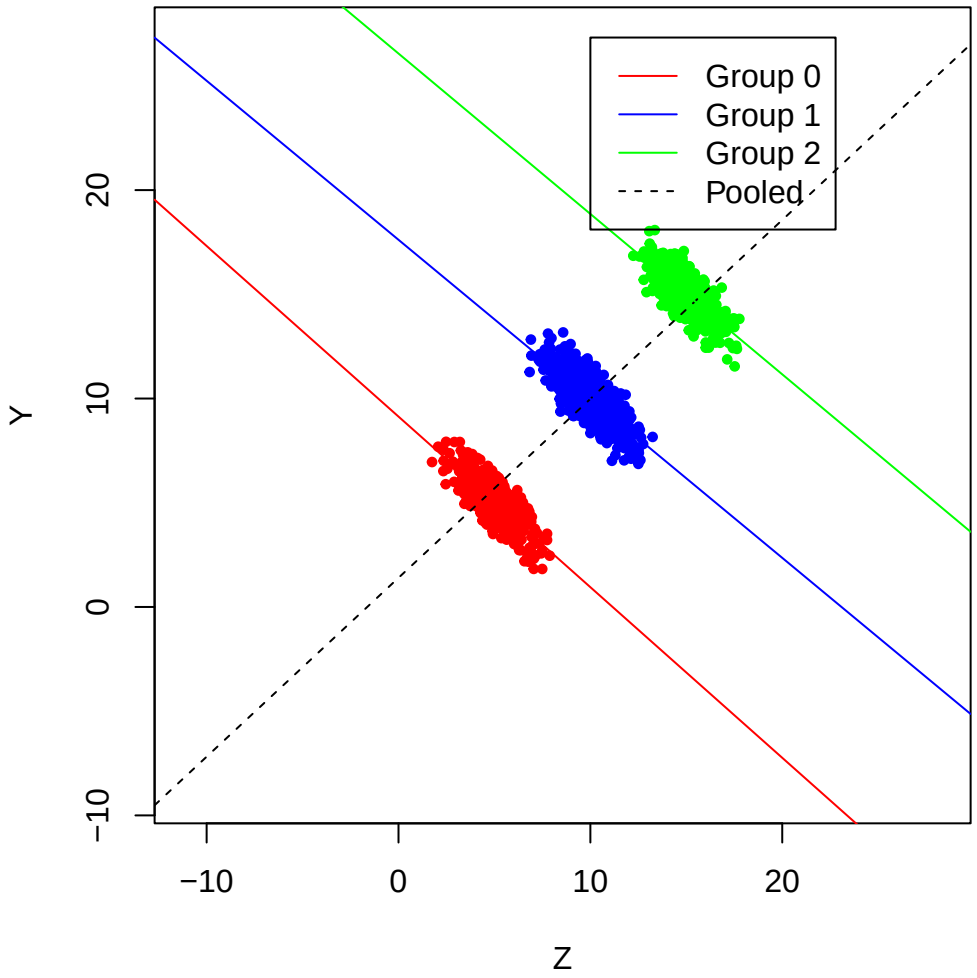
```
Y1<-Y[X>4.5 & X < 5.5];Z1<-Z[X>4.5 & X < 5.5];       #First subset analysis
Y2<-Y[X>9.5 & X < 10.5];Z2<-Z[X>9.5 & X < 10.5];     #Second subset analysis
Y3<-Y[X>14.5 & X < 15.5];Z3<-Z[X>14.5 & X < 15.5];   #Third subset analysis
par(mar=c(4,2,1,0),pty='s',mfrow=c(2,2))             #Set up the plotting margins
plot(Z1,Y1,pch=19,cex=0.6)
plot(Z2,Y2,pch=19,cex=0.6)
plot(Z3,Y3,pch=19,cex=0.6)
```



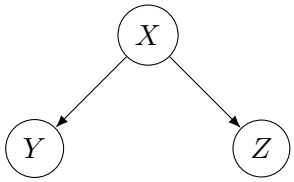
In each subset, we see the negative correlation from the original conditional calculation. We can check this by regressing Y on Z for each subset.

```
coef(summary(lm(Y1~Z1)))                          #Group 1 regression

+               Estimate Std. Error   t value      Pr(>|t|)
+ (Intercept)  9.1336287 0.14557132  62.74333 4.733761e-236
+ Z1          -0.8184506 0.02868607 -28.53129 3.872360e-106

coef(summary(lm(Y2~Z2)))                          #Group 2 regression

+               Estimate Std. Error   t value      Pr(>|t|)
+ (Intercept) 17.6160738 0.23954157  73.54078  0.000000e+00
+ Z2          -0.7629346 0.02374499 -32.13033 3.096694e-146

coef(summary(lm(Y3~Z3)))                          #Group 3 regression

+               Estimate Std. Error   t value      Pr(>|t|)
+ (Intercept) 26.5516171 0.47620626  55.75655 4.768122e-208
+ Z3          -0.7695225 0.03163317 -24.32644  6.308926e-85

coef(summary(lm(c(Y1,Y2,Y3)~c(Z1,Z2,Z3))))     #Pooled regression

+                Estimate Std. Error  t value     Pr(>|t|)
+ (Intercept)   1.3976736 0.12496988 11.18408 4.216643e-28
+ c(Z1, Z2, Z3) 0.8583356 0.01172109 73.23004 0.000000e+00
```

```
par(mar=c(4,4,1,0),pty='s')
plot(Z,Y,type='n')
points(Z1,Y1,pch=19,cex=0.6,col='red')
points(Z2,Y2,pch=19,cex=0.6,col='blue')
points(Z3,Y3,pch=19,cex=0.6,col='green')
abline(lm(c(Y1,Y2,Y3)~c(Z1,Z2,Z3)),lty=2)
abline(lm(Y1~Z1),col='red')
abline(lm(Y2~Z2),col='blue')
abline(lm(Y3~Z3),col='green')
legend(10,max(Y),c('Group 0','Group 1','Group 2','Pooled'),
                 col=c('red','blue','green','black'),lty=c(1,1,1,2))
```



We can now repeat the analysis, but assuming, as per the graph,

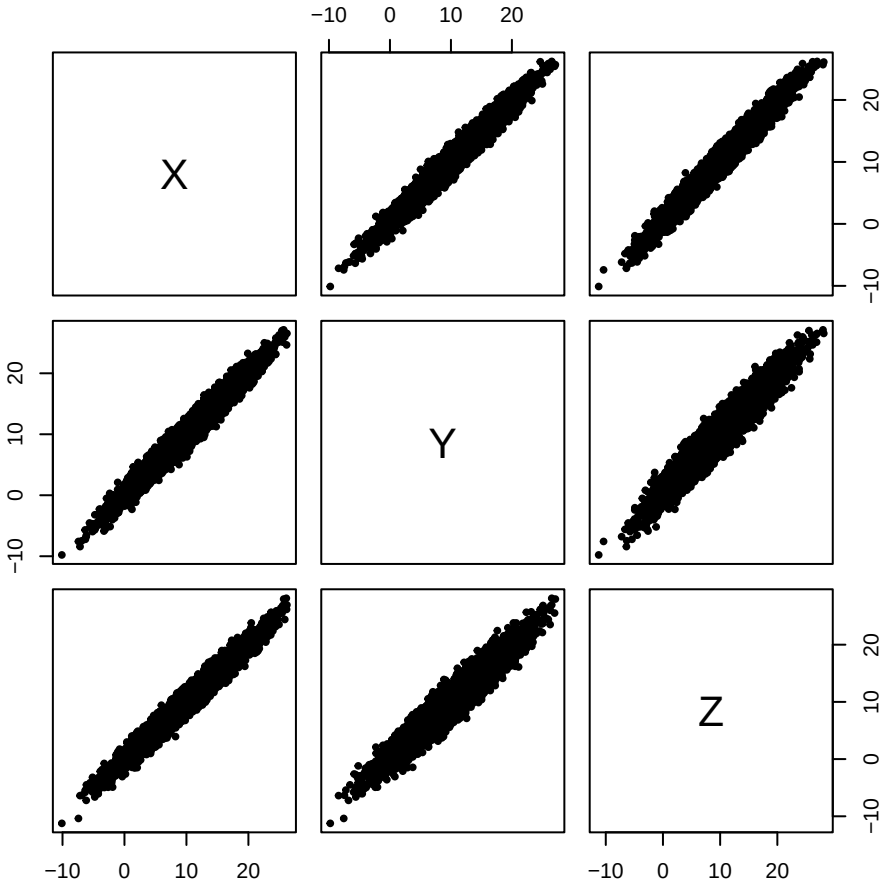


that Y and Z are **conditionally independent** given X.

```
set.seed(2111)                          #Set the random number generator seed value
n<-10000                                #Set the sample size
X<-rnorm(n,10,5)                        #Generate the X random variables
Sig.YZ<-matrix(c(1,0,0,1),2,2)          #Conditional independence for Y,Z given X
YZ<-mvrnorm(n,mu=c(0,0),Sigma=Sig.YZ)   #Generate the Y,Z variables
Y<-X+YZ[,1];Z<-X+YZ[,2]                 #Change the mean according to X
cor(cbind(X,Y,Z))                       #Compute the unconditional correlation

+           X         Y         Z
+ X 1.0000000 0.9809467 0.9806338
+ Y 0.9809467 1.0000000 0.9616723
+ Z 0.9806338 0.9616723 1.0000000
```

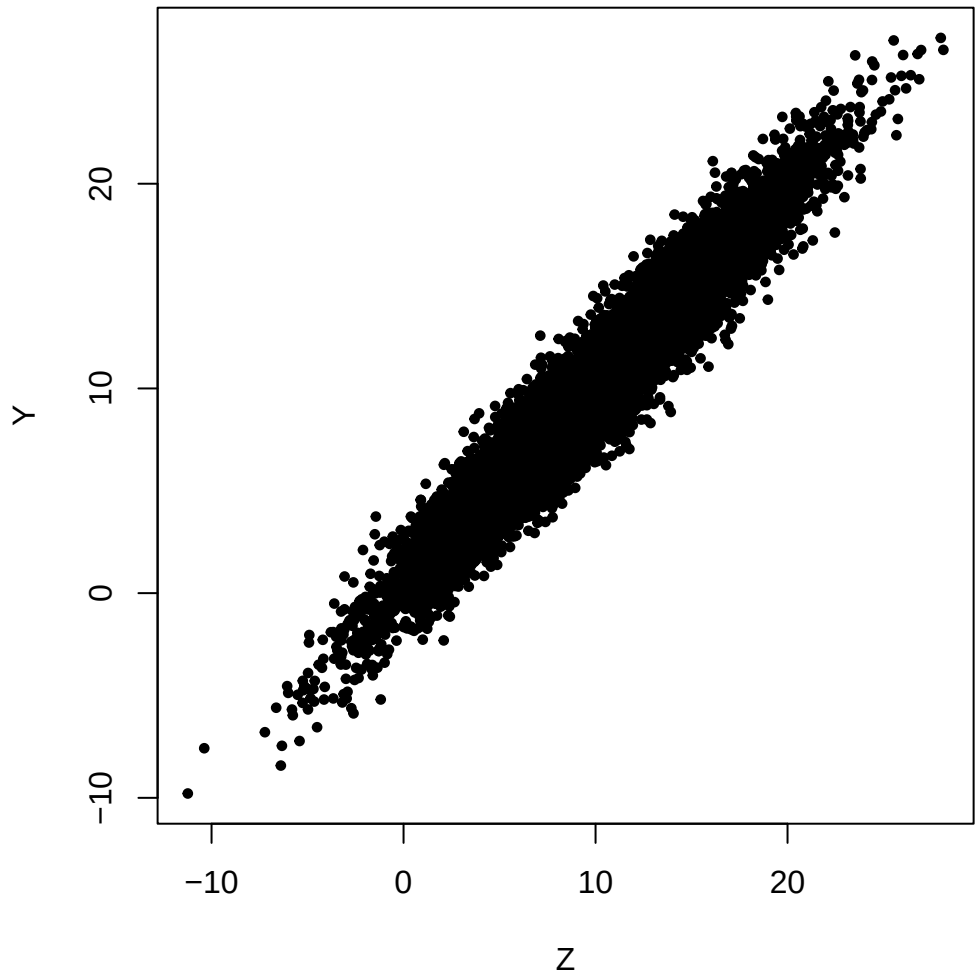
We see that marginally all the variables are positively correlated.

```
par(mar=c(4,4,1,0),pty='s')             #Set up the plotting margins
pairs(cbind(X,Y,Z),pch=19,cex=0.6)      #Scatterplot matrix
```



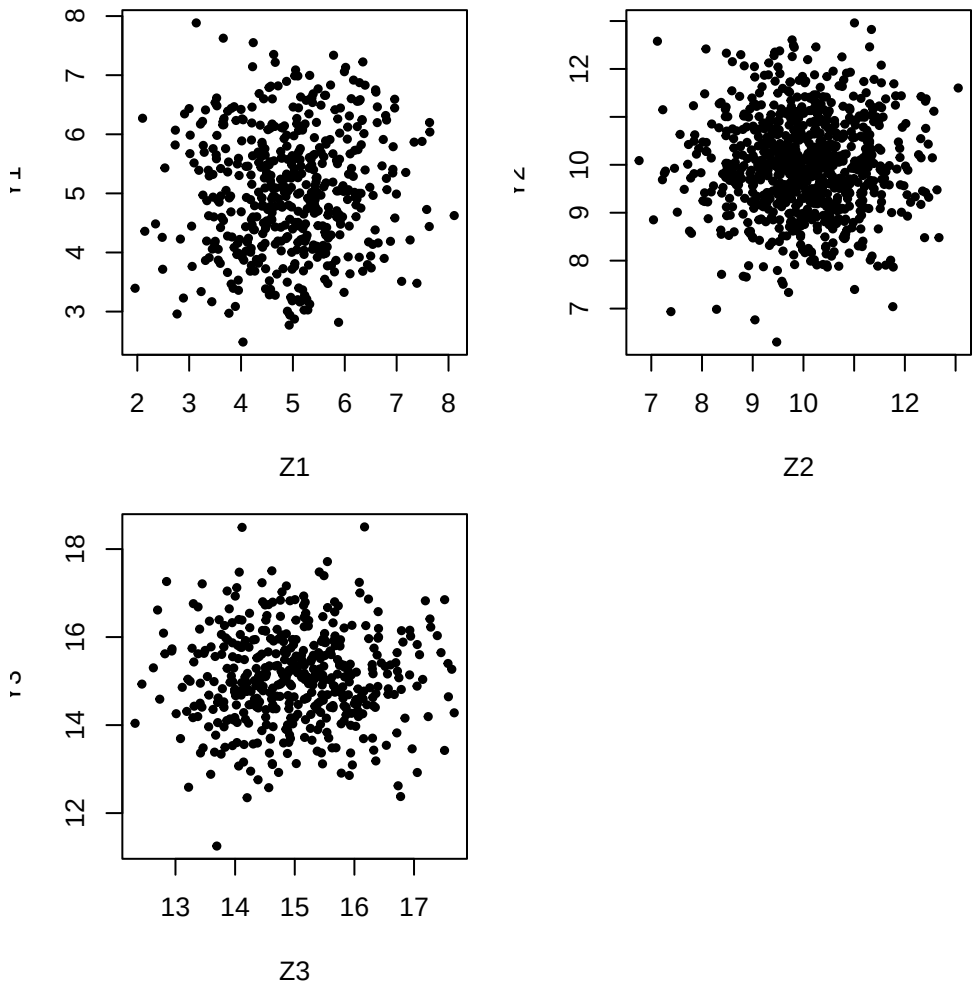
There is evidently strong positive correlation between each pair of variables, including Y and Z.

```
par(mar=c(4,4,1,0),pty='s')             #Set up the plotting margins
plot(Z,Y,pch=19,cex=0.6)
```



We now examine three subsets defined by different ranges of X:

```
Y1<-Y[X>4.5 & X < 5.5];Z1<-Z[X>4.5 & X < 5.5];       #First subset analysis
Y2<-Y[X>9.5 & X < 10.5];Z2<-Z[X>9.5 & X < 10.5];     #Second subset analysis
Y3<-Y[X>14.5 & X < 15.5];Z3<-Z[X>14.5 & X < 15.5];   #Third subset analysis
par(mar=c(4,2,1,0),pty='s',mfrow=c(2,2))             #Set up the plotting margins
plot(Z1,Y1,pch=19,cex=0.6)
plot(Z2,Y2,pch=19,cex=0.6)
plot(Z3,Y3,pch=19,cex=0.6)
```

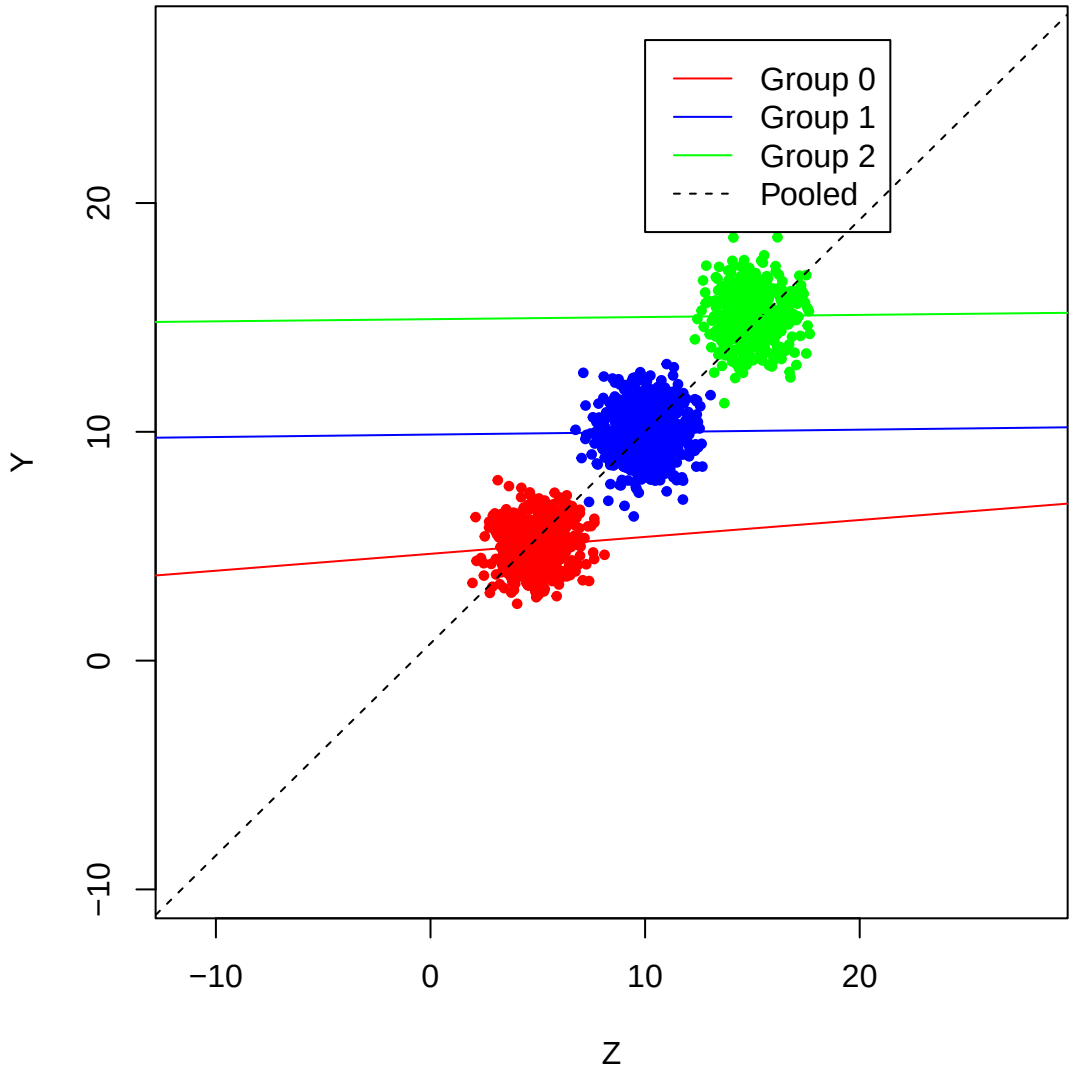


In each subset, we see the negligible correlation from the original conditional calculation. We can check this by regressing Y on Z for each subset.

```
coef(summary(lm(Y1~Z1)))                        #Group 1 regression

+               Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) 4.66768535 0.22422265 20.817189 2.180142e-69
+ Z1          0.07387431 0.04424949  1.669495 9.565927e-02

coef(summary(lm(Y2~Z2)))                        #Group 2 regression

+               Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept) 9.87950842 0.35272687 28.0089478 6.30107e-121
+ Z2          0.01069578 0.03494438  0.3060802  7.59623e-01

coef(summary(lm(Y3~Z3)))                        #Group 3 regression

+                Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept) 14.922338303 0.70900368 21.0469123 1.424837e-69
+ Z3           0.009389903 0.04714401  0.1991749 8.422130e-01

coef(summary(lm(c(Y1,Y2,Y3)~c(Z1,Z2,Z3))))     #Pooled regression

+                Estimate  Std. Error    t value     Pr(>|t|)
+ (Intercept)   0.7461296 0.094180708   7.922319 4.097475e-15
+ c(Z1, Z2, Z3) 0.9262747 0.008836666 104.821734 0.000000e+00
```

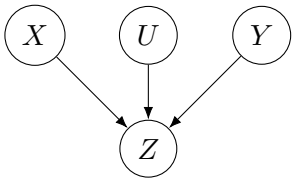
The slope coefficient is essentially zero in each case, confirming conditional independence, and yet marginally Y and Z are strongly positively correlated,

```
par(mar=c(4,4,1,0),pty='s')
plot(Z,Y,type='n')
points(Z1,Y1,pch=19,cex=0.6,col='red')
points(Z2,Y2,pch=19,cex=0.6,col='blue')
points(Z3,Y3,pch=19,cex=0.6,col='green')
abline(lm(c(Y1,Y2,Y3)~c(Z1,Z2,Z3)),lty=2)
abline(lm(Y1~Z1),col='red')
abline(lm(Y2~Z2),col='blue')
abline(lm(Y3~Z3),col='green')
legend(10,max(Y),c('Group 0','Group 1','Group 2','Pooled'),
                 col=c('red','blue','green','black'),lty=c(1,1,1,2))
```



2. COLLIDERS

The following graph

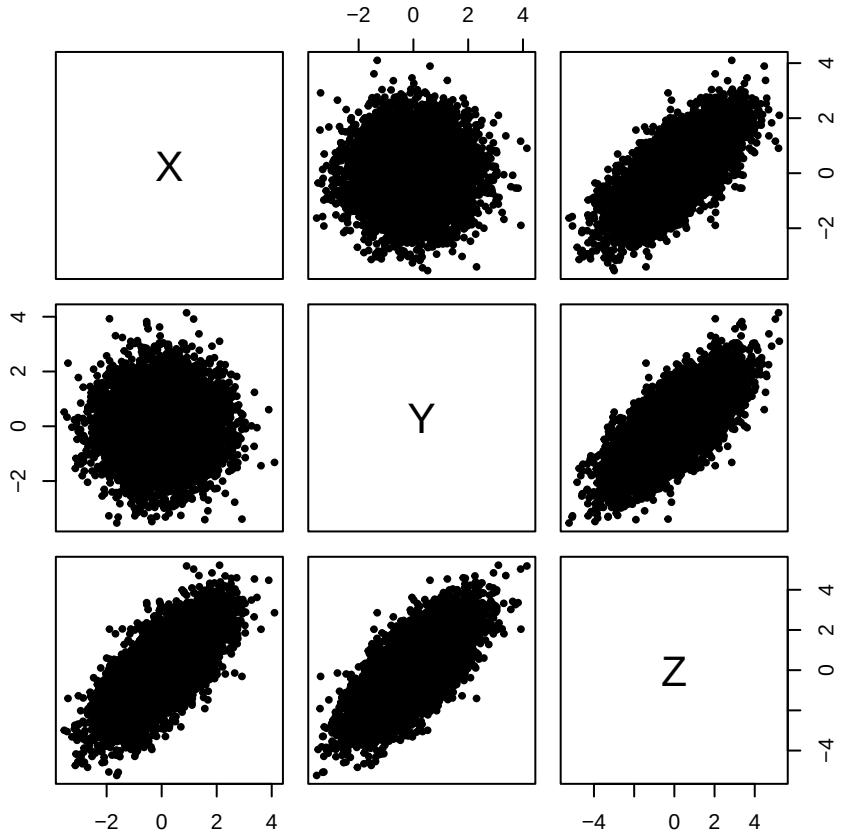


encodes the dependencies between the four random variables (X, Y, Z, U): the joint density can be represented

$$f_{X,Y,Z,U}(x, y, z, u) = f_X(x) f_Y(y) f_U(u) f_{Z|X,Y,U}(z|x, y, u)$$

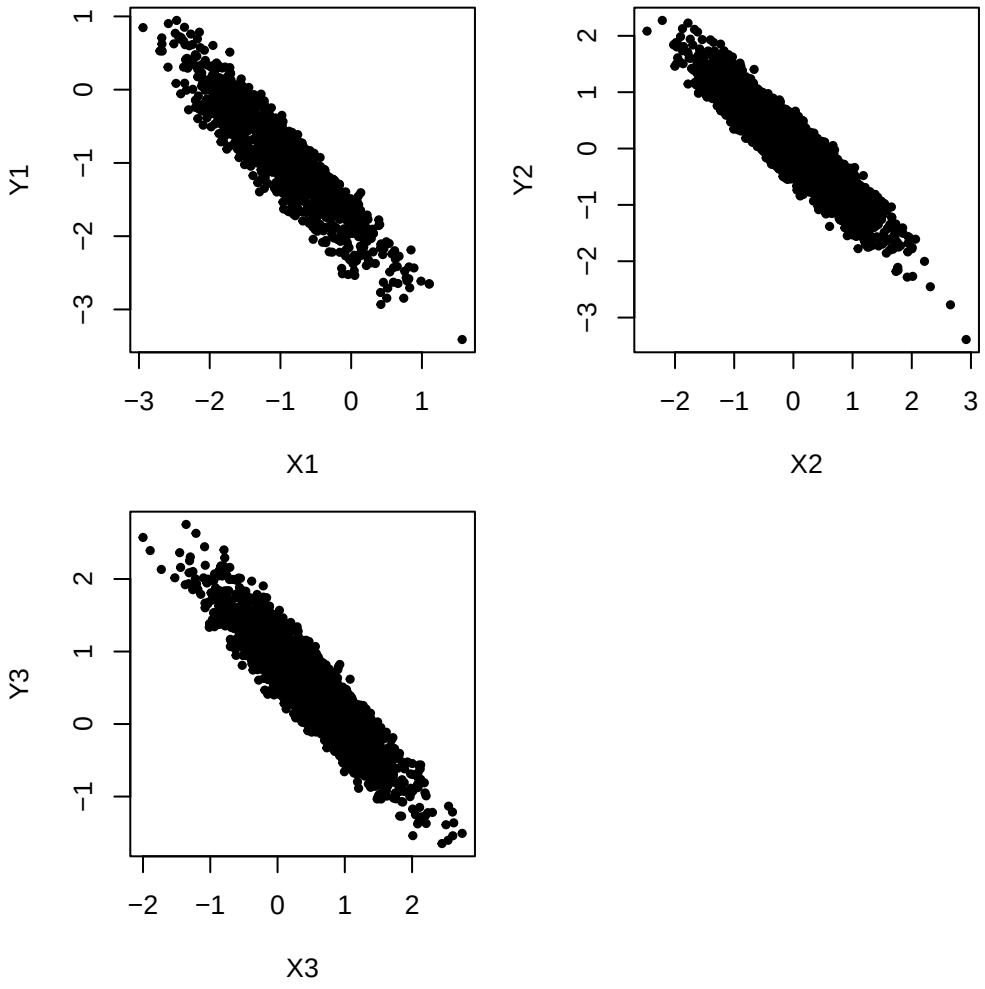
that is, X, Y, U are mutually independent. Suppose that X and Y are distributed as standard Normal random variables, $U \sim \text{Normal}(0, 0.1^2)$, and $Z = X + Y + U$.

```
set.seed(2101)                          #Set the random number generator seed value
n<-10000                                #Set the sample size
X<-rnorm(n,0,1)                         #Generate the X random variables
Y<-rnorm(n,0,1)                         #Generate the Y random variables
U<-rnorm(n,0,0.1)                       #Generate the U random variables
Z<-X+Y+U
par(mar=c(3,2,1,0))                     #Set up the plotting margins
pairs(cbind(X,Y,Z),pch=19,cex=0.6)
```

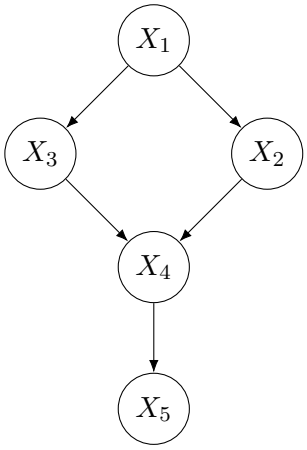


If we condition on the value of collider Z, and inspect the joint density of X and Y given Z, we see that X and Y are conditionally **dependent**.

```
par(mar=c(3,2,1,0))                                    #Set up the plotting margins
X1<-X[Z>-2.5 & Z < -1.5];Y1<-Y[Z>-2.5 & Z < -1.5];     #First subset analysis
X2<-X[Z>-0.5 & Z < 0.5];Y2<-Y[Z>-0.5 & Z < 0.5];       #Second subset analysis
X3<-X[Z>0.5 & Z < 1.5];Y3<-Y[Z>0.5 & Z < 1.5];         #Third subset analysis
par(mar=c(4,3,1,0),pty='s',mfrow=c(2,2))               #Set up the plotting margins
plot(X1,Y1,pch=19,cex=0.6)
plot(X2,Y2,pch=19,cex=0.6)
plot(X3,Y3,pch=19,cex=0.6)
cor(X1,Y1)

+ [1] -0.9142475

cor(X2,Y2)

+ [1] -0.9100298

cor(X3,Y3)

+ [1] -0.9109185
```



Example 1: Consider the graph



The implied probability distribution is

$$f_{X_1,X_2,X_3,X_4,X_5}(x_1,x_2,x_3,x_4,x_5) = f_{X_1}(x_1)f_{X_2|X_1}(x_2|x_1)f_{X_3|X_1}(x_3|x_1)f_{X_4|X_2,X_3}(x_4|x_2,x_3)f_{X_5|X_4}(x_5|x_4)$$

We have that there are two paths connecting X_2 and X_3, namely

$$(X_2, X_1, X_3) \qquad (X_2, X_4, X_3)$$

- the set $S \equiv \{X_1\}$ d-separates X_2 and X_3 as conditioning on X_1 blocks the path (X_2, X_1, X_3); we do not need to consider the second path here, as that is already blocked by the collider X_4;
- the set $S \equiv \{X_1, X_4\}$ does not d-separate X_2 and X_3 as conditioning on the collider X_4 opens the path;
- the set $S \equiv \{X_1, X_5\}$ does not d-separate X_2 and X_3 as conditioning on X_5, a descendant of X_4, opens the path at the collider.

To illustrate these results, we perform a simulation:

```
set.seed(43)
n<-1000
X1<-rnorm(n,1,1)
X2<-rnorm(n,X1,1)
X3<-rnorm(n,X1,1)
X4<-rnorm(n,X2+X3,1)
X5<-rnorm(n,X4,1)
cor(cbind(X1,X2,X3,X4,X5))
```

```
+           X1        X2        X3        X4        X5
+ X1 1.0000000 0.7179843 0.7177349 0.7754169 0.7373380
+ X2 0.7179843 1.0000000 0.4939446 0.8167909 0.7651028
+ X3 0.7177349 0.4939446 1.0000000 0.7965510 0.7530419
+ X4 0.7754169 0.8167909 0.7965510 1.0000000 0.9395833
+ X5 0.7373380 0.7651028 0.7530419 0.9395833 1.0000000
```

We can check the d-separation here using regression. We condition on the relevant variables by including them as predictors.

- $S \equiv \{X_1\}$ d-separates X_2 and X_3

```
summary(lm(X2~X3))$coef

+              Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) 0.4601408 0.04992427  9.216776 1.765162e-19
+ X3          0.5147179 0.02868085 17.946393 1.244543e-62

summary(lm(X3~X2))$coef

+              Estimate Std. Error  t value     Pr(>|t|)
+ (Intercept) 0.5744852 0.04647570 12.36098 9.253159e-33
+ X2          0.4740096 0.02641253 17.94639 1.244543e-62
```

The two variables X_2 and X_3 are dependent, as the slope coefficient is non-zero. We now include X_1 as a predictor.

```
summary(lm(X2~X3+X1))$coef

+                Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) -0.07357452 0.04585679 -1.604441 1.089335e-01
+ X3          -0.04594547 0.03295795 -1.394063 1.636091e-01
+ X1           1.09503001 0.04620065 23.701612 7.439007e-99

summary(lm(X3~X2+X1))$coef

+                Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept)  0.01478124 0.04407664  0.3353532 7.374292e-01
+ X2          -0.04234298 0.03037379 -1.3940634 1.636091e-01
+ X1           1.05053469 0.04436882 23.6773161 1.075811e-98
```

In each case, after including X_1, the coefficient of the other regressor becomes close to zero.

- the set $S \equiv \{X_1, X_4\}$ does not d-separate X_2 and X_3

```
summary(lm(X2~X3+X1+X4))$coef

+                Estimate Std. Error     t value      Pr(>|t|)
+ (Intercept) -0.01558546 0.03112569  -0.5007267  6.166741e-01
+ X3          -0.54465543 0.02665949 -20.4300766  9.321772e-78
+ X1           0.50167819 0.03578069  14.0209213  6.811246e-41
+ X4           0.52281106 0.01525553  34.2702558 1.166823e-170

summary(lm(X3~X2+X1+X4))$coef

+                Estimate Std. Error    t value      Pr(>|t|)
+ (Intercept)  0.02483552 0.03104929   0.799874  4.239745e-01
+ X2          -0.54219655 0.02653913 -20.430077  9.321772e-78
+ X1           0.51994317 0.03542036  14.679216  2.614779e-44
+ X4           0.50358274 0.01581951  31.833022 5.798631e-154
```

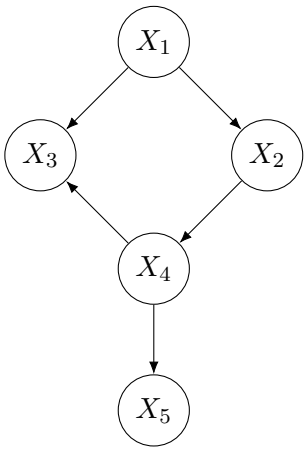
After including X_1 and X_4, the coefficient of X_3 in the first analysis, and X_2 in the second analysis, becomes non-zero.

- the set $S \equiv \{X_1, X_5\}$ does not d-separate X_2 and X_3.

```
summary(lm(X2~X3+X1+X5))$coef

+                Estimate Std. Error     t value     Pr(>|t|)
+ (Intercept) -0.009686452 0.03699708  -0.2618167 7.935169e-01
+ X3          -0.380599759 0.03014958 -12.6237178 5.375796e-34
+ X1           0.682446683 0.04116576  16.5780167 1.074329e-54
+ X5           0.346937155 0.01487318  23.3263533 2.285559e-96

summary(lm(X3~X2+X1+X5))$coef

+               Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept)  0.04571273 0.03607389   1.267197 2.053810e-01
+ X2          -0.36240099 0.02870794 -12.623718 5.375796e-34
+ X1           0.67761978 0.03997277  16.952035 7.940014e-57
+ X5           0.32809458 0.01475215  22.240452 2.724060e-89
```

After including X_1 and X_5, the coefficient of X_33 in the first analysis, and X_2 in the second analysis, becomes non-zero.

Example 2: Consider the graph



The implied probability distribution is

$$f_{X_1,X_2,X_3,X_4,X_5}(x_1,x_2,x_3,x_4,x_5) = f_{X_1}(x_1)f_{X_2|X_1}(x_2|x_1)f_{X_4|X_2}(x_4|x_2)f_{X_3|X_1,X_4}(x_3|x_1,x_4)f_{X_5|X_4}(x_5|x_4)$$

The same two paths connecting X_2 and X_3 are present, but

- the set $S \equiv \{X_1\}$ does not d-separate X_2 and X_3 as conditioning on X_1 blocks the path (X_2, X_1, X_3); but there is an open path through X_4;
- the set $S \equiv \{X_1, X_4\}$ d-separates X_2 and X_3 as conditioning on X_1 and X_4 blocks both paths;
- the set $S \equiv \{X_1, X_5\}$ does not d-separate X_2 and X_3 as the second path remains open at X_4.

We can check this in simulation:

```
X1<-rnorm(n,1,1)
X2<-rnorm(n,X1,1)
X4<-rnorm(n,X2,1)
X3<-rnorm(n,X1+X4,1)
X5<-rnorm(n,X4,1)
```

Again, we can check for d-separation in this case using regression.

- $S \equiv \{X_1\}$ does not d-separate X_2 and X_3

```
summary(lm(X2~X3+X1))$coef

+               Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) -0.04131217 0.03702937 -1.115659 2.648367e-01
+ X3           0.35268288 0.01521919 23.173565 2.203979e-95
+ X1           0.31004897 0.04111230  7.541513 1.043267e-13

summary(lm(X3~X2+X1))$coef

+               Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept) -0.01996549 0.06215680 -0.3212116 7.481173e-01
+ X2           0.99259472 0.04283306 23.1735646 2.203979e-95
+ X1           1.05363328 0.06256884 16.8395844 3.443879e-56
```

  In each case, after including X_1, the coefficient of the other regressor is non-zero.

- the set $S \equiv \{X_1, X_4\}$ d-separates X_2 and X_3

```
summary(lm(X2~X3+X1+X4))$coef

+               Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept) -0.04428818 0.03250915 -1.3623296 1.734019e-01
+ X3           0.01851411 0.02353282  0.7867357 4.316237e-01
+ X1           0.47994434 0.03741281 12.8283442 5.646272e-35
+ X4           0.50247755 0.02912892 17.2501276 1.522789e-58

summary(lm(X3~X2+X1+X4))$coef

+               Estimate Std. Error    t value      Pr(>|t|)
+ (Intercept) -0.03476823 0.04378585 -0.7940516  4.273546e-01
+ X2           0.03354484 0.04263800  0.7867357  4.316237e-01
+ X1           1.00302508 0.04410230 22.7431493  1.487590e-92
+ X4           1.00145445 0.03145961 31.8330223 5.798631e-154
```

  After including X_1 and X_4, the coefficient of X_3 in the first analysis, and X_2 in the second analysis, becomes zero.

- the set $S \equiv \{X_1, X_5\}$ does not d-separate X_2 and X_3.

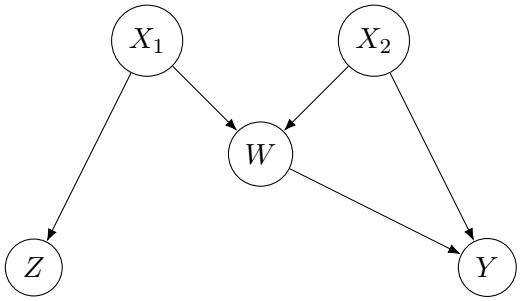
```
summary(lm(X2~X3+X1+X5))$coef
```

```
+                Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept) -0.02916628 0.03559760 -0.8193327 4.127926e-01
+ X3           0.23222662 0.01964361 11.8219910 2.851977e-30
+ X1           0.35959996 0.03986236  9.0210404 9.376853e-19
+ X5           0.18104355 0.01971766  9.1817958 2.390699e-19

summary(lm(X3~X2+X1+X5))$coef

+                Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) 0.009319753 0.05378931  0.173264 8.624791e-01
+ X2          0.529886531 0.04482211 11.821991 2.851977e-30
+ X1          0.987953903 0.05424034 18.214377 3.323363e-64
+ X5          0.491945695 0.02681838 18.343604 5.638032e-65
```

After including X_1 and X_5, the coefficient of X_33 in the first analysis, and X_2 in the second analysis, becomes non-zero.

In general, when considering d-separation of two variables X and Y by a set of variables, S, we need to consider **all paths** between X and Y, not merely those involving S - however, in the first example X_4 is a collider so we do not need to consider the second path when taking $S \equiv \{X_1\}$.

Example 3: Consider the DAG



which implies the structural decomposition

$$f_{X_1}(x_1)f_{X_2}(x_2)f_{W|X_1,X_2}(w|x_1,x_2)f_{Z|X_1}(z|x_1)f_{Y|X_2,W}(y|x_2,w).$$

For example, we might have

$$\begin{aligned} X_1 &\sim Normal(1,1) \\ X_2 &\sim Normal(2,1) \\ W|X_1 = x_1, X_2 = x_2 &\sim Normal(x_1 + x_2, 1) \\ Z|X_1 = x_1 &\sim Normal(2 + 2x_1, 1) \\ Y|X_2 = x_2, W = w &\sim Normal(x_2 + w, 1) \end{aligned}$$

In this DAG, there are no directed paths from Z to Y; thus, Z is **not** a cause of Y. If we unpick the structural specifications, we have either that

$$Y = X_2 + W + \epsilon$$

or

$$Y = X_2 + (X_1 + X_2) + \varepsilon = X_1 + 2X_2 + \varepsilon.$$

Thus, manipulating Z by intervention has no impact on Y. There are, however, two undirected paths

$$\text{Path 1:} \quad Z \to X_1 \to W \to X_2 \to Y$$

$$\text{Path 2:} \quad Z \to X_1 \to W \to Y.$$

which have the potential to introduce bias into estimation of the (zero) causal effect.

```
set.seed(43)
n<-1000
X1<-rnorm(n,1,1)
X2<-rnorm(n,2,1)
W<-rnorm(n,X1+X2,1)
Z<-rnorm(n,2+2*X1,1)
Y<-rnorm(n,X2+W,1)
cor(cbind(X1,X2,W,Z,Y))
```

```
+           X1         X2         W          Z         Y
+ X1 1.00000000 0.04794271 0.6078559 0.90328569 0.4346003
+ X2 0.04794271 1.00000000 0.5895617 0.05901935 0.7536782
+ W  0.60785594 0.58956173 1.0000000 0.55824686 0.8815832
+ Z  0.90328569 0.05901935 0.5582469 1.00000000 0.4058155
+ Y  0.43460031 0.75367824 0.8815832 0.40581553 1.0000000
```

If we attempt an uncorrected analysis, and simply regress Y on Z, the results imply that Z does influence Y.

```
summary(lm(Y~Z))$coef
```

```
+              Estimate Std. Error  t value     Pr(>|t|)
+ (Intercept) 3.0349189 0.15948952 19.02895 4.004626e-69
+ Z           0.4833063 0.03445509 14.02714 6.233441e-41
```

Therefore, we need to perform some form of adjustment to block the undirected paths by conditioning. Now W is a collider on Path 1 from Z to Y, so this path is blocked. Conditioning on W *opens* the confounding path. Therefore $Z \perp\!\!\!\perp Y$ (as there is no open path between them), but

$$Z \not\perp\!\!\!\perp Y \mid W$$

Unconditional on W, the effect of Z on Y is confounded by the backdoor path Path 2. Conditioning on W alone opens Path 1, therefore to block both paths need to condition on

$$S \equiv \{W, X_2\} \qquad \text{or} \qquad S = \{X_1\}.$$

We consider five models

$$M_0 \quad Y = Z + \epsilon$$

$$M_1 \quad Y = Z + W + \epsilon$$

$$M_2 \quad Y = Z + W + X_2 + \epsilon$$

$$M_3 \quad Y = Z + X_1 + \epsilon$$

$$M_4 \quad Y = Z + W + X_1 + X_2 + \epsilon$$

Models M_2, M_3 and M_4 should yield unbiased estimators of the (null) effect of Z on Y.

```
summary(lm(Y~Z))$coef             #M0 - uncorrected

+              Estimate Std. Error  t value     Pr(>|t|)
+ (Intercept) 3.0349189 0.15948952 19.02895 4.004626e-69
+ Z           0.4833063 0.03445509 14.02714 6.233441e-41

summary(lm(Y~Z+W))$coef           #M1 - conditioned on W

+               Estimate Std. Error   t value      Pr(>|t|)
+ (Intercept)  1.1391572 0.08767980 12.992241  9.064370e-36
+ Z           -0.1493541 0.02093105 -7.135527  1.853246e-12
+ W            1.4745925 0.02723446 54.144370 3.865321e-299

summary(lm(Y~Z+W+X2))$coef        #M2 - conditioned on (W,X2)

+                Estimate Std. Error    t value      Pr(>|t|)
+ (Intercept) -0.07767336 0.08861465 -0.8765296  3.809536e-01
+ Z            0.02173540 0.01851196  1.1741269  2.406249e-01
+ W            1.01584028 0.02976870 34.1244448 1.158489e-169
+ X2           0.96724624 0.04218159 22.9305330  8.920611e-94

summary(lm(Y~Z+X1))$coef          #M3 - conditioned on X1

+               Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) 3.62787847 0.18989154 19.105004 1.410294e-69
+ Z           0.08570895 0.07912919  1.083152 2.790028e-01
+ X1          0.99395864 0.17868442  5.562649 3.412874e-08

summary(lm(Y~Z+W+X1+X2))$coef     #M4 - conditioned on (W,X1,X2)

+                 Estimate Std. Error    t value      Pr(>|t|)
+ (Intercept) -0.059234282 0.09766668 -0.6064943  5.443249e-01
+ Z            0.009558079 0.03279712  0.2914305  7.707829e-01
+ W            1.009915593 0.03256260 31.0145833 2.616934e-148
+ X1           0.036408761 0.08093055  0.4498766  6.528974e-01
+ X2           0.973080915 0.04414658 22.0420462  5.388882e-88
```

We now run a simulation study, replicating 1000 times for $n = 500$.

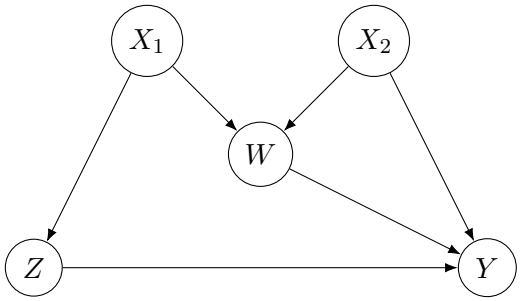
```
nreps<-1000
n<-500
ests<-matrix(0,nrow=nreps,ncol=5)
for(irep in 1:nreps){
    X1<-rnorm(n,1,1)
    X2<-rnorm(n,2,1)
    W<-rnorm(n,X1+X2,1)
    Z<-rnorm(n,2+2*X1,1)
    Y<-rnorm(n,X2+W,1)
    ests[irep,1]<-coef(lm(Y~Z))[2]              #M0 - uncorrected
    ests[irep,2]<-coef(lm(Y~Z+W))[2]            #M1 - conditioned on W
    ests[irep,3]<-coef(lm(Y~Z+W+X2))[2]         #M2 - conditioned on (W,X2)
    ests[irep,4]<-coef(lm(Y~Z+X1))[2]           #M3 - conditioned on X1
    ests[irep,5]<-coef(lm(Y~Z+W+X1+X2))[2]      #M4 - conditioned on (W,X1,X2)
}
Bias<-apply(ests,2,mean)*sqrt(n)
```

```
MSE<-apply(ests^2,2,mean)*n
names(Bias)<-names(MSE)<-c('M0','M1','M2','M3','M4')
Bias                                          #Bias

+          M0          M1          M2          M3          M4
+  8.94936146 -4.07347810 -0.00261349  0.01984241 -0.02285131

MSE                                           #Mean-square error

+         M0         M1         M2         M3         M4
+ 81.2418380 17.0053131  0.3211222  5.6236450  1.1346711
```

Therefore we have the following results:

Model		$\sqrt{n} \times$ Bias	$n \times$ MSE
M_0	$Y = Z + \epsilon$	8.9494	81.2418
M_1	$Y = Z + W + \epsilon$	-4.0735	17.0053
M_2	$Y = Z + W + X_2 + \epsilon$	-0.0026	0.3211
M_3	$Y = Z + X_1 + \epsilon$	0.0198	5.6236
M_4	$Y = Z + W + X_1 + X_2 + \epsilon$	-0.0229	1.1347

Example 4: Consider the modified DAG



which implies the structural decomposition

$$f_{X_1}(x_1)f_{X_2}(x_2)f_{W|X_1,X_2}(w|x_1,x_2)f_{Z|X_1}(z|x_1)f_{Y|X_2,W,Z}(y|x_2,w,z).$$

For example, we might have

$$X_1 \sim Normal(1,1)$$
$$X_2 \sim Normal(2,1)$$
$$W|X_1 = x_1, X_2 = x_2 \sim Normal(x_1 + x_2, 1)$$
$$Z|X_1 = x_1 \sim Normal(2 + 2x_1, 1)$$
$$Y|X_2 = x_2, W = w \sim Normal(x_2 + w + z, 1)$$

In this DAG, there is one directed path from Z to Y; thus, Z **is** a cause of Y. From the structural specifications, we have either that

$$Y = X_2 + W + Z + \epsilon$$

or

$$Y = X_2 + (X_1 + X_2) + Z + \varepsilon = X_1 + 2X_2 + Z + \varepsilon.$$

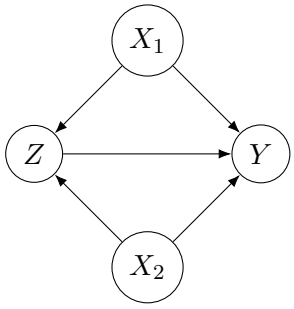
so the impact of manipulating Z by intervention is that changing Z by one unit changes the expected value of Y by one unit. The same confounding paths are present. We consider again the five models; models M_2, M_3 and M_4 should yield unbiased estimators of the effect of Z on Y, which is the coefficient 1.

```
nreps<-1000
n<-500
ests<-matrix(0,nrow=nreps,ncol=5)
for(irep in 1:nreps){
    X1<-rnorm(n,1,1)
    X2<-rnorm(n,2,1)
    W<-rnorm(n,X1+X2,1)
    Z<-rnorm(n,2+2*X1,1)
    Y<-rnorm(n,X2+W+Z,1)
    ests[irep,1]<-coef(lm(Y~Z))[2]              #M0 - uncorrected
    ests[irep,2]<-coef(lm(Y~Z+W))[2]            #M1 - conditioned on W
    ests[irep,3]<-coef(lm(Y~Z+W+X2))[2]         #M2 - conditioned on (W,X2)
    ests[irep,4]<-coef(lm(Y~Z+X1))[2]           #M3 - conditioned on X1
    ests[irep,5]<-coef(lm(Y~Z+W+X1+X2))[2]      #M4 - conditioned on (W,X1,X2)
}
Bias<-apply(ests-1,2,mean)*sqrt(n)
MSE<-apply((ests-1)^2,2,mean)*n
names(Bias)<-names(MSE)<-c('M0','M1','M2','M3','M4')
Bias                                            #Bias

+          M0          M1          M2          M3          M4
+  9.00187653 -4.03001620  0.02396948  0.11182724  0.03956394

MSE                                             #Mean-square error

+         M0         M1          M2         M3         M4
+ 82.3537793 16.6632231  0.3238196  6.0576331  1.0788292
```

We have the following results:

Model		$\sqrt{n} \times$ Bias	$n \times$ MSE
M_0	$Y = Z + \epsilon$	9.0019	82.3538
M_1	$Y = Z + W + \epsilon$	-4.0300	16.6632
M_2	$Y = Z + W + X_2 + \epsilon$	0.0240	0.3238
M_3	$Y = Z + X_1 + \epsilon$	0.1118	6.0576
M_4	$Y = Z + W + X_1 + X_2 + \epsilon$	0.0396	1.0788

Example 5: Consider the following DAG



which implies the structural decomposition

$$f_{X_1}(x_1)f_{X_2}(x_2)f_{Z|X_1,X_2}(z|x_1,x_2)f_{Y|X_1,X_2,Z}(y|x_1,x_2,z).$$

Suppose that

$$X_1 \sim Normal(1,1)$$
$$X_2 \sim Normal(2,1)$$
$$Z|X_1 = x_1, X_2 = x_2 \sim Normal(x_1 + x_2, 1)$$
$$Y|X_1 = x_1, X_2 = x_2, Z = z \sim Normal(x_1 - x_2 + z, 1)$$

In this DAG, there is one directed path from Z to Y; thus, Z **is** a cause of Y. From the structural specifications, we have either that

$$Y = X_1 - X_2 + Z + \epsilon$$

so the impact of manipulating Z by intervention is that changing Z by one unit changes the expected value of Y by one unit, but also

$$Y = X_1 - X_2 + (X_1 + X_2) + \varepsilon = 2X_1 + \varepsilon.$$

There are two confounding paths

$$\text{Path 1:} \quad Z \to X_1 \to Y$$
$$\text{Path 2:} \quad Z \to X_2 \to Y.$$

We consider four models

$$M_0 \quad Y = Z + \epsilon$$
$$M_1 \quad Y = Z + X_1 + \epsilon$$
$$M_2 \quad Y = Z + X_2 + \epsilon$$
$$M_3 \quad Y = Z + X_1 + X_2 + \epsilon$$

In a simulation study

```
nreps<-1000
n<-500
ests<-matrix(0,nrow=nreps,ncol=4)
for(irep in 1:nreps){
    X1<-rnorm(n,1,1)
```

```
    X2<-rnorm(n,2,1)
    Z<-rnorm(n,X1+X2,1)
    Y<-rnorm(n,X1-X2+Z,1)
    ests[irep,1]<-coef(lm(Y~Z))[2]              #M0 - uncorrected
    ests[irep,2]<-coef(lm(Y~Z+X1))[2]           #M1 - conditioned on X1
    ests[irep,3]<-coef(lm(Y~Z+X2))[2]           #M2 - conditioned on X2
    ests[irep,4]<-coef(lm(Y~Z+X1+X2))[2]        #M3 - conditioned on (X1,X2)
}
Bias<-apply(ests-1,2,mean)*sqrt(n)
MSE<-apply((ests-1)^2,2,mean)*n
names(Bias)<-names(MSE)<-c('M0','M1','M2','M3')
```

We have the following results:

Model		$\sqrt{n} \times$ Bias	$n \times$ MSE
M_0	$Y = Z + \epsilon$	-0.0685	1.0502
M_1	$Y = Z + X_1 + \epsilon$	-11.2339	126.9481
M_2	$Y = Z + X_2 + \epsilon$	11.1569	125.2333
M_3	$Y = Z + X_1 + X_2 + \epsilon$	-0.0282	1.0907

In this case, the **unadjusted** model also produces an unbiased estimator of the causal effect of Z on Y. This is because of the **specific** data generating process chosen. Specifically, in the outcome model, the random variables $X_1 - X_2$ and Z are **uncorrelated**:

$$\begin{aligned}
\text{Cov}[(X_1 - X_2), Z] &= \text{Cov}[(X_1 - X_2), (X_1 + X_2 + \varepsilon)] && \varepsilon \perp\!\!\!\perp (X_1, X_2) \\
&= \text{Cov}[(X_1 - X_2), (X_1 + X_2)] + \text{Cov}[(X_1 - X_2), \varepsilon] \\
&= \left\{\mathbb{E}[(X_1^2 - X_2^2)] - \mathbb{E}[(X_1 - X_2)]\mathbb{E}[(X_1 + X_2)]\right\} + 0 && \text{zero by independence} \\
&= \{(1 + 1) - (4 + 1)) - (1 - 2)(1 + 2)\} = -3 + 3 = 0
\end{aligned}$$

as if $X \sim Normal(\mu, \sigma^2)$, $\mathbb{E}[X^2] = \mu^2 + \sigma^2$. If the simulation model is varied, M_0 **no longer** produces an unbiased result.

```
ests<-matrix(0,nrow=nreps,ncol=4)
for(irep in 1:nreps){
    X1<-rnorm(n,1,1)
    X2<-rnorm(n,2,1)
    Z<-rnorm(n,X1+2*X2,1)                       #Model changed here !
    Y<-rnorm(n,X1-X2+Z,1)
    ests[irep,1]<-coef(lm(Y~Z))[2]              #M0 - uncorrected
    ests[irep,2]<-coef(lm(Y~Z+X1))[2]           #M1 - conditioned on X1
    ests[irep,3]<-coef(lm(Y~Z+X2))[2]           #M2 - conditioned on X2
    ests[irep,4]<-coef(lm(Y~Z+X1+X2))[2]        #M3 - conditioned on (X1,X2)
}
Bias<-apply(ests-1,2,mean)*sqrt(n)
MSE<-apply((ests-1)^2,2,mean)*n
names(Bias)<-names(MSE)<-c('M0','M1','M2','M3')
```

Here, from the structural specifications, we have again that

$$Y = X_1 - X_2 + Z + \epsilon$$

Model		$\sqrt{n} \times$ Bias	$n \times$ MSE
M_0	$Y = Z + \epsilon$	-3.7725	14.7228
M_1	$Y = Z + X_1 + \epsilon$	-8.9756	80.8179
M_2	$Y = Z + X_2 + \epsilon$	11.1468	124.9455
M_3	$Y = Z + X_1 + X_2 + \epsilon$	-0.0847	0.9767

so the impact of manipulating Z by intervention is that changing Z by one unit changes the expected value of Y by one unit, but now

$$Y = X_1 - X_2 + (X_1 + 2X_2) + \varepsilon = 2X_1 + X_2 + \varepsilon.$$

4. Selection Bias

The vignette for `R` package `ggdag` contains two examples on selection bias

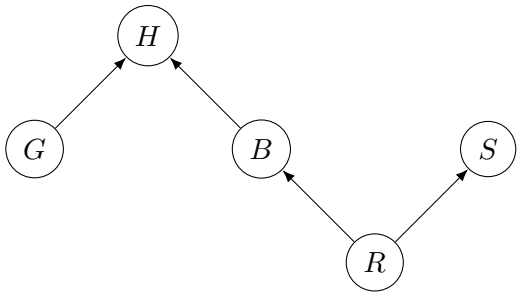
`https://cran.r-project.org/web/packages/ggdag/vignettes/bias-structures.html`

that inspire the numerical examples below.

Example 1: Consider the DAG for binary variables

- S – smoking status
- G – glioma
- H – hospitalization
- B – broken bone
- R – tendency for reckless behaviour

We wish to study the association between S and G in a hospitalized cohort.



The corresponding probability model factorizes as

$$f_{R,B,S,G,H}(r, b, s, g, h) = f_R(r) f_G(g) f_{B|R}(b|r) f_{S|R}(s|r) f_{H|B,G}(h|b, g)$$

In this graph, we have one path from S to G

- this is an undirected path,
- the path is blocked at collider H.

Therefore $G \perp\!\!\!\perp S$, but conditioning on H renders G and S **dependent**. That is, if we carry out a study in the general population, no association between will be detected, however, if we only look in hospitalized subjects, an association will be detected.

```
set.seed(2384)
n<-1000000                          #Large population
R<-rbinom(n,1,0.2)                  #Simulate R
G<-rbinom(n,1,0.02)                 #Simulate G
pS<-c(0.05,0.25)                    #Simulate S, with different probs for R=0 and R=1
S<-rbinom(n,1,pS[R+1])
pB<-c(0.01,0.10)                    #Simulate B, with different probs for R=0 and R=1
B<-rbinom(n,1,pB[R+1])
pH<-c(0.01,0.9,0.05,0.95)           #Simulate H, with different probs for different values
H<-rbinom(n,1,pH[B+G+B*G+1])        #of the quantity B+G+G*B
cor(G,S)                            #No association in general population

+ [1] -0.0006665788

cor(G[H==1],S[H==1])                #Negative correlation in hospitalized individuals

+ [1] -0.1065372

cor(G[H==0],S[H==0])                #No appreciable association in non hospitalized

+ [1] 0.001231331
```

We can explore the same results using regression:

```
my.data<-data.frame(R,G,S,B,H)
summary(lm(G~S,data=my.data))$coef                  #Analysis in general population

+                  Estimate   Std. Error     t value  Pr(>|t|)
+ (Intercept)  0.0203025106 0.0001477228 137.4365419 0.0000000
+ S           -0.0003285636 0.0004929107  -0.6665783 0.5050417
```

From the analysis in the general population, ignoring hospitalization status, there is no association between Smoking and Glioma. However, performing an analysis stratifying by hospitalization status yields different results.

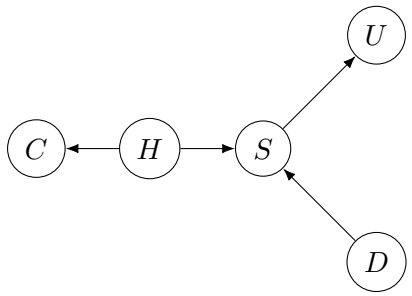
```
summary(lm(G~S,subset=(H==1),data=my.data))$coef    #Analysis in hospitalized

+               Estimate  Std. Error  t value      Pr(>|t|)
+ (Intercept)  0.3706017 0.002236389 165.7143  0.000000e+00
+ S           -0.1474371 0.006023914 -24.4753 1.499987e-131

summary(lm(G~S,subset=(H==0),data=my.data))$coef    #Analysis in non-hospitalized

+                 Estimate   Std. Error   t value  Pr(>|t|)
+ (Intercept) 0.0020873931 4.927136e-05 42.365239 0.0000000
+ S           0.0002000502 1.668789e-04  1.198774 0.2306161
```

Thus from the analysis of hospitalized subjects, it seems that Smoking is protective for Glioma.

Example 2: Consider the DAG for variables

- C – CD4 count (continuous)
- D – assignment of new HIV drug (binary)
- H – underlying HIV severity (binary)
- S – symptoms (binary)
- U – follow up indicator (binary, $U = 1$ implies that participant stays in trial)

We wish to examine the impact of D on C by studying patients who stay in the trial. Suppose the relevant DAG is



The corresponding probability model factorizes as

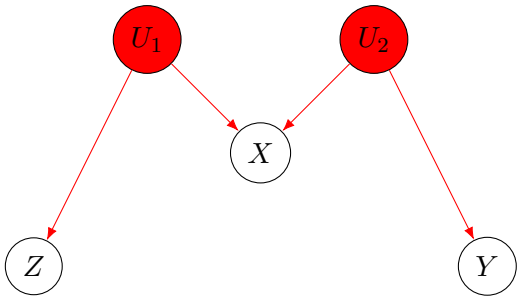
$$f_{H,D,C,S,U}(h,d,c,s,u) = f_H(h)f_D(d)f_{C|H}(c|h)f_{S|H,D}(s|h,d)f_{U|S}(u|s)$$

In this graph, we have one path from D to C: this is an undirected path, but the path is blocked at collider S. Therefore $D \perp\!\!\!\perp C$. However, conditioning on U, which is a descendant of S, renders D and C **dependent**. Therefore, regressing C on D in the subgroup where $U = 1$ (that is, individuals who are followed in the trial) will indicate an association.

```
set.seed(2384)
n<-1000000                          #Large population
H<-rbinom(n,1,0.2)                  #Simulate H
D<-rbinom(n,1,0.5)                  #Simulate D
C<-rnorm(n,100-20*H,10)             #Simulate C
pS<-c(0.01,0.9,0.05,0.95)           #Simulate S, with different probs for different values
S<-rbinom(n,1,pS[H+D+H*D+1])        #of the quantity H+D+H*D
pU<-c(0.9,0.3)                      #Simulate U
U<-rbinom(n,1,pU[S+1])
coef(summary(lm(C~D)))

+                Estimate Std. Error      t value  Pr(>|t|)
+ (Intercept) 95.97691408 0.01810726 5300.4650599 0.0000000
+ D            0.02425598 0.02559712    0.9476058 0.3433304

coef(summary(lm(C~D, subset=(U==1))))

+              Estimate Std. Error    t value Pr(>|t|)
+ (Intercept) 98.152659 0.01896969 5174.18351        0
+ D           -1.922286 0.03400687  -56.52641        0

coef(summary(lm(C~D, subset=(U==0))))

+              Estimate Std. Error  t value Pr(>|t|)
+ (Intercept) 87.944580 0.04032272 2181.018        0
+ D            7.930516 0.04649965  170.550        0
```

5. M-BIAS

Consider the DAG with two unmeasured confounders:



We have that X, Y and Z are *independent*; the (true but hidden) path between Z and Y is blocked at collider X.

```
set.seed(43)
n<-1000
U1<-rnorm(n,1,1)
U2<-rnorm(n,2,1)
X<-rnorm(n,U1+U2,1)
Z<-rnorm(n,2+2*U1,1)
Y<-rnorm(n,3-U2,1)
cor(cbind(U1,U2,X,Z,Y))
```

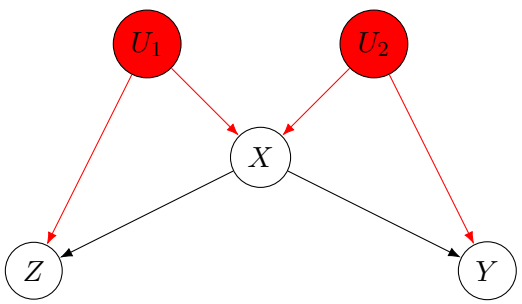
```
+            U1          U2          X           Z            Y
+ U1 1.00000000  0.04794271  0.6078559 0.903285693  0.011284669
+ U2 0.04794271  1.00000000  0.5895617 0.059019354 -0.716871840
+ X  0.60785594  0.58956173  1.0000000 0.558246862 -0.392340308
+ Z  0.90328569  0.05901935  0.5582469 1.000000000  0.001654136
+ Y  0.01128467 -0.71687184 -0.3923403 0.001654136  1.000000000
```

```
summary(lm(Y~Z))$coef
```

```
+               Estimate Std. Error     t value     Pr(>|t|)
+ (Intercept) 1.013726740 0.09342831 10.85031668 5.263348e-26
+ Z           0.001054719 0.02018365  0.05225612 9.583351e-01
```

It is evident that Z and Y are not associated; they are uncorrelated, and the estimated coefficient of Z in the regression of Y on Z is close to zero.

In an analysis, however, suppose we condition on X: the relevant DAG is presented below



In the modelled DAG, $Y \perp\!\!\!\perp Z \mid X$; however, conditioning on X *opens* the *hidden* path through U_1 and U_2, so there is now an open biasing path.

```
summary(lm(Y~Z+X))$coef

+               Estimate Std. Error    t value     Pr(>|t|)
+ (Intercept)  1.6230853 0.08977596  18.079286 2.070535e-63
+ Z            0.2044120 0.02143145   9.537947 1.076994e-20
+ X           -0.4739813 0.02788555 -16.997379 4.296672e-57
```

Inclusion of X in the regression induces bias into the estimation.