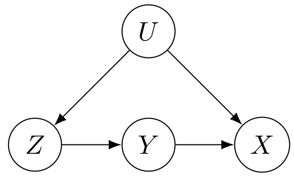
# Backdoor path with a collider and unmeasured confounding




The corresponding probability model factorizes as

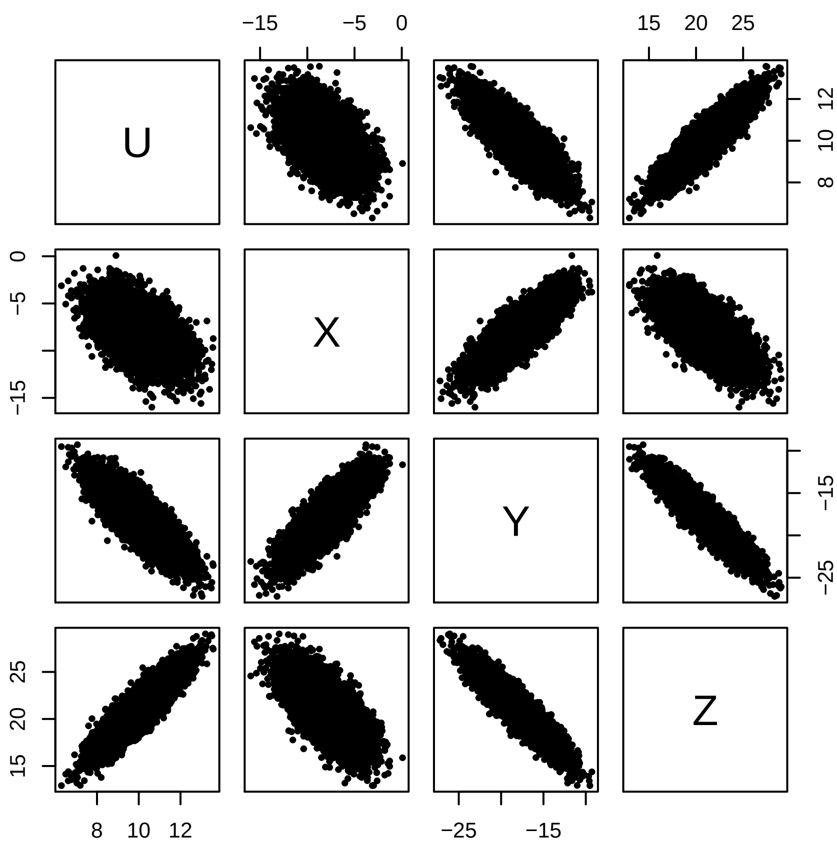
$$f_{U,X,Y,Z}(u,x,y,z) = f_U(u) f_{Z|U}(z|u) f_{Y|Z}(y|z) f_{X|U,Y}(x|u,y)$$

In this graph, we have two paths from $Z$ to $Y$

- Path $(Z,Y)$ this is a directed path;
- Path $(Z,U,X,Y)$: this is an undirected path that is also a backdoor path.

However the second path is blocked at the collider $X$, so there is no open backdoor path, and thus the effect of $Z$ on $Y$ is only found through the first.

```
set.seed(2384)
n<-10000
U<-rnorm(n,10,1)
Z<-rnorm(n,2*U+1,1)
Y<-rnorm(n,-Z+3,1)
X<-rnorm(n,Y+U,1)
data1<-data.frame(U,X,Y,Z);pairs(data1,pch=19,cex=0.5)
```




If we regress $Y$ on $Z$, then the correct relationship is recovered.

```r
round(coef(summary(lm(Y~Z))),6)

+               Estimate Std. Error    t value Pr(>|t|)
+ (Intercept)   2.961994   0.094022   31.50325        0
+ Z            -0.998035   0.004448 -224.39271        0
```

However, if we condition on $X$ in the regression, we see that bias is introduced in the estimation of the coefficient.

```r
round(coef(summary(lm(Y~Z+X))),6)

+               Estimate Std. Error    t value Pr(>|t|)
+ (Intercept)   0.903804   0.072623   12.44514        0
+ Z            -0.727314   0.004393 -165.54528        0
+ X             0.453505   0.004917   92.23862        0
```

If we condition on $U$ only, then the direct effect of $Z$ on $Y$ is correctly captured, as the path is still blocked at $X$

```r
round(coef(summary(lm(Y~Z+U))),6)

+               Estimate Std. Error     t value Pr(>|t|)
+ (Intercept)   2.932047   0.100382   29.208980  0.00000
+ Z            -1.005670   0.010007 -100.498340  0.00000
+ U             0.019045   0.022360    0.851756  0.39437
```

If we condition on $U$ and $X$, then the direct effect of $Z$ on $Y$ is also not captured.

```r
round(coef(summary(lm(Y~Z+X+U))),6)

+               Estimate Std. Error   t value Pr(>|t|)
+ (Intercept)   1.474144   0.072171  20.42563        0
+ Z            -0.502210   0.008640 -58.12750        0
+ X             0.498937   0.004952 100.76459        0
+ U            -0.493843   0.016552 -29.83663        0
```

This is due to *selection bias*: conditioning on a descendant of $Y$ will lead to bias in most circumstances. Consider the simple chain graph The corresponding probability model factorizes as

$$Z \longrightarrow Y \longrightarrow X$$

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

Clearly we can integrate out $x$ from the joint density to leave

$$f_{Y,Z}(y,z) = f_Z(z) f_{Y|Z}(y|z)$$

leaving the $(Z,Y)$ relationship unchanged. However, we have that

$$f_{Y|X,Z}(y|x,z) = \frac{f_{X,Y,Z}(x,y,z)}{f_{X,Z}(x,z)} = \frac{f_Z(z) f_{Y|Z}(y|z) f_{X|Y,Z}(x|y,z)}{f_Z(z) f_{X|Z}(x|z)} = \frac{f_{X|Y,Z}(x|y,z)}{f_{X|Z}(x|z)} f_{Y|Z}(y|z)$$

and in general

$$\frac{f_{X|Y,Z}(x|y,z)}{f_{X|Z}(x|z)} \neq 1$$

so

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

Notice, however, that we can change the data generating model to make the analyses agree. If the conditional model for $Y$ given $Z$ is instead

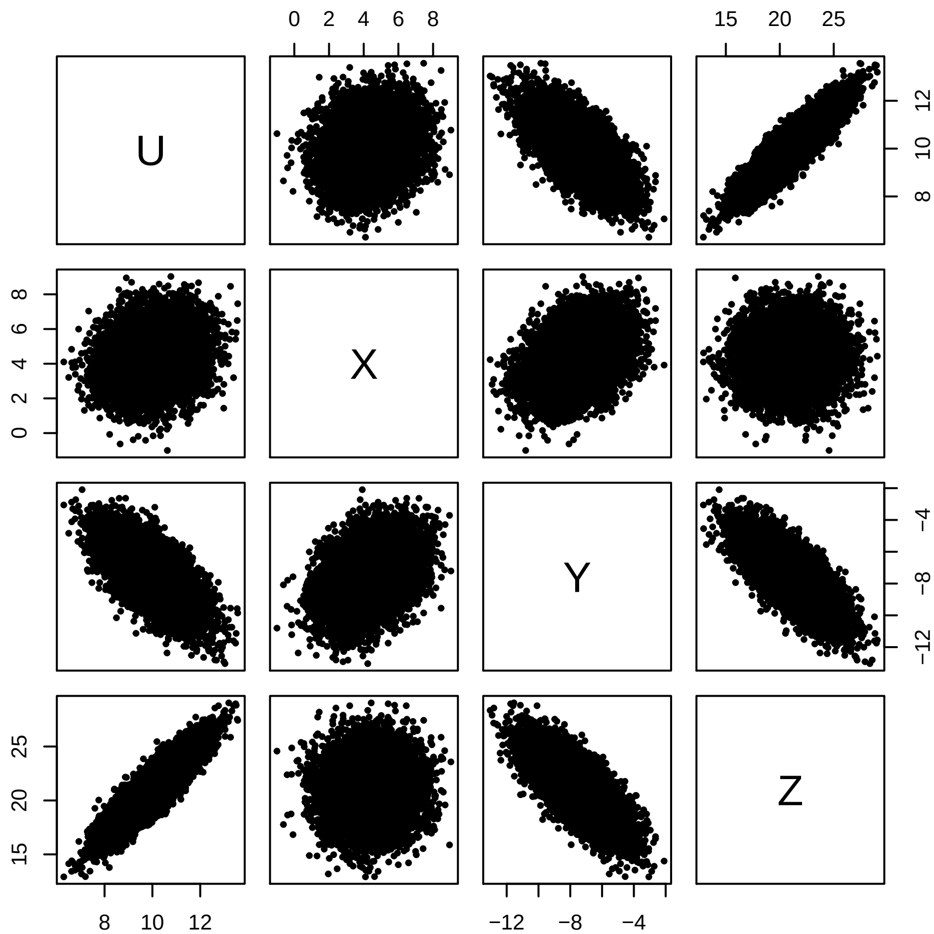
$$Y = -0.5Z + 3 + \epsilon$$

and then

$$X = 0.75Y + U + \varepsilon$$

then the effect of conditioning changes.

```
set.seed(2384)
n<-10000
U<-rnorm(n,10,1)
Z<-rnorm(n,2*U+1,1)
Y<-rnorm(n,-0.5*Z+3,1)
X<-rnorm(n,0.75*Y+U,1)
data2<-data.frame(U,X,Y,Z);pairs(data2,pch=19,cex=0.5)
```




If we regress $Y$ on $Z$, then the **correct** relationship is recovered.

```
round(coef(summary(lm(Y~Z))),6)

+              Estimate Std. Error    t value Pr(>|t|)
+ (Intercept)  2.961994   0.094022   31.50325        0
+ Z           -0.498035   0.004448 -111.97548        0
```

Now, if we condition on $X$ in the regression, we see that bias is still **not present**, even though there is an open, biasing path.

```
round(coef(summary(lm(Y~Z+X))),6)

+              Estimate Std. Error    t value Pr(>|t|)
+ (Intercept)  1.342382   0.080650   16.64451        0
+ Z           -0.509786   0.003656 -139.42652        0
+ X            0.426448   0.006137   69.49304        0
```

If we condition on $U$ only, then the direct effect of $Z$ on $Y$ is **correctly captured**, as the path is still blocked at $X$

```
round(coef(summary(lm(Y~Z+U))),6)

+              Estimate Std. Error    t value Pr(>|t|)
+ (Intercept)  2.932047   0.100382  29.208980  0.00000
+ Z           -0.505670   0.010007 -50.532496  0.00000
+ U            0.019045   0.022360   0.851756  0.39437
```

However, if we condition on $U$ and $X$, then the direct effect of $Z$ on $Y$ is **not** captured.

```
round(coef(summary(lm(Y~Z+X+U))),6)

+              Estimate Std. Error   t value Pr(>|t|)
+ (Intercept)  1.880322   0.081181  23.16219        0
+ Z           -0.321823   0.008334 -38.61466        0
+ X            0.480458   0.006337  75.82198        0
+ U           -0.472560   0.018960 -24.92430        0
```

As a final summary, we can inspect the **inverse** of the sample correlation matrices:

```
round(solve(cor(data1)),6)

+           U         X         Y         Z
+ U  6.088401 -2.031254  2.461387 -4.562629
+ X -2.031254  4.020629 -4.965237 -0.029805
+ Y  2.461387 -4.965237 12.168425  5.592981
+ Z -4.562629 -0.029805  5.592981 10.175987
```

```
round(solve(cor(data2)),6)

+           U         X         Y         Z
+ U  6.088401 -1.354396  1.124682 -4.552405
+ X -1.354396  1.787544 -1.522344 -0.004926
+ Y  1.124682 -1.522344  3.550755  1.711425
+ Z -4.552405 -0.004926  1.711425  6.354784
```

Note that in both cases, the entry in the position relating $X$ and $Z$ is almost zero. This is an indication that conditional on the other variables, $X$ and $Z$ are uncorrelated (actually independent here in this Gaussian case). This entry corresponds to the **partial correlation** between the two variables.