XYZ

Spurious association

Suppose that $Y \perp Z$ and $Y \leftarrow X \rightarrow Z$. In particular, consider the linear model

$$Y = \beta_x X + \beta_z Z + \varepsilon$$

where $\beta_x = 1$ and $\beta_z = 0$. If we consider the submodel

$$Y = \tilde{\beta}_x X + \tilde{\varepsilon}$$

then it may be that $\tilde{\beta}_x \neq 0$.

```
rm(list=ls())
library(MASS)
n = 200
set.seed(2)

x = rnorm(n)
z = x + rnorm(n,0,0.5)
X = cbind(x,z)
beta = c(1,0)
y = X %*% beta + rnorm(n)
p=2

M = 1
X_M = X[,M]
P_M = X_M %*% solve(t(X_M)%*% X_M) %*% t(X_M)
ginvX_M = ginv(X_M)
beta_M = rep(0,p)
beta_M[M] = ginvX_M %*% P_M %*% X %*% beta
rbind(beta,round(beta_M,2))
```

```
FALSE      [,1] [,2]
FALSE beta    1    0
FALSE         1    0
```

```
round(summary(lm(y ~ 0+X_M))$coef,4)
```

```
FALSE     Estimate Std. Error t value Pr(>|t|)
FALSE X_M   1.0517      0.067 15.7068        0
```

```
round(summary(lm(y ~ 0+X))$coef,4)
```

```
FALSE    Estimate Std. Error t value Pr(>|t|)
FALSE Xx   0.9283     0.1567  5.9250   0.0000
FALSE Xz   0.1238     0.1422  0.8708   0.3849
```

```
M = 2
X_M = X[,M]
P_M = X_M %*% solve(t(X_M)%*% X_M) %*% t(X_M)
ginvX_M = ginv(X_M)
beta_M = rep(0,p)
beta_M[M] = ginvX_M %*% P_M %*% X %*% beta
rbind(beta,round(beta_M,2))
```

```
FALSE      [,1] [,2]
FALSE beta    1 0.00
FALSE         0 0.82
```

```
round(summary(lm(y ~ 0+X_M))$coef,4)
```

```
FALSE     Estimate Std. Error t value Pr(>|t|)
FALSE X_M   0.8853     0.0658 13.4548        0
```

```
round(summary(lm(y ~ 0+X))$coef,4)
```

```
FALSE    Estimate Std. Error t value Pr(>|t|)
FALSE Xx   0.9283     0.1567  5.9250   0.0000
FALSE Xz   0.1238     0.1422  0.8708   0.3849
```

Another example

```
rm(list=ls())

n = 100
p = 3
beta = c(-1,2,0)

set.seed(2)
rho = 0.9
R = matrix(rho,ncol=p,nrow=p) + diag(rep(1-rho,p))
X = mvrnorm(n, mu=rep(0,p), Sigma=R)

y = X%*%beta + rnorm(n)

M = c(1,2)
X_M = X[,M]
P_M = X_M %*% solve(t(X_M)%*% X_M) %*% t(X_M)
ginvX_M = ginv(X_M)
beta_M = rep(0,p)
beta_M[M] = ginvX_M %*% P_M %*% X %*% beta
rbind(beta,round(beta_M,2))
```

```
FALSE      [,1] [,2] [,3]
FALSE beta   -1    2    0
FALSE        -1    2    0
```

```
M = c(1,3)
X_M = X[,M]
P_M = X_M %*% solve(t(X_M)%*% X_M) %*% t(X_M)
ginvX_M = ginv(X_M)
beta_M = rep(0,p)
beta_M[M] = ginvX_M %*% P_M %*% X %*% beta
rbind(beta,round(beta_M,2))
```

```
FALSE       [,1] [,2] [,3]
FALSE beta -1.00    2 0.00
FALSE       0.04    0 0.92
```

```
round(summary(lm(y ~ 0+X_M))$coef,4)
```

```
FALSE      Estimate Std. Error t value Pr(>|t|)
FALSE X_M1   0.2943     0.2646  1.1123   0.2687
FALSE X_M2   0.6651     0.2598  2.5596   0.0120
```

```r
round(summary(lm(y ~ 0+X))$coef,4)
```

```
FALSE    Estimate Std. Error t value Pr(>|t|)
FALSE X1  -0.7104     0.2476 -2.8689   0.0051
FALSE X2   1.9397     0.2534  7.6553   0.0000
FALSE X3  -0.2308     0.2371 -0.9733   0.3328
```