Problem 3

```
set.seed(123)
( K <- cbind(c(10,7,7,0),c(7,20,0,7),c(7,0,30,7),c(0,7,7,40)) )
```

```
##      [,1] [,2] [,3] [,4]
## [1,]   10    7    7    0
## [2,]    7   20    0    7
## [3,]    7    0   30    7
## [4,]    0    7    7   40
```

```
data <- as.data.frame(mvrnorm(n=10000,mu=c(0,0,0,0),Sigma=solve(K)))

colnames(data) <- c("X1","X2","X3","X4")
```

Conditional independency

It represents following independencies:

$X_1 \perp\!\!\!\perp X_4 | X_2, X_3$ and $X_2 \perp\!\!\!\perp X_3 | X_1, X_4$

The corresponding graph

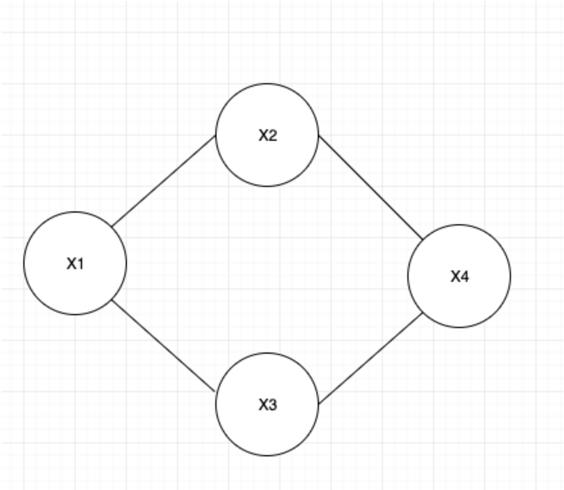


Figure 1: fig 3-1

Fit with OLS

```
lmodel = lm(X1 ~ X4 + X2 + X3, data=data)
summary(lmodel)
```

```
##
## Call:
## lm(formula = X1 ~ X4 + X2 + X3, data = data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.36729 -0.21127  0.00304  0.21389  1.20994
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001934   0.003141   0.616    0.538
## X4           0.007927   0.020037   0.396    0.692
## X2          -0.682729   0.012203 -55.950   <2e-16 ***
## X3          -0.695282   0.015540 -44.741   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3141 on 9996 degrees of freedom
## Multiple R-squared:  0.4564, Adjusted R-squared:  0.4563
## F-statistic:  2798 on 3 and 9996 DF,  p-value: < 2.2e-16
```

X4 is not significant while X2 and X3 are. This means X4 and X1 is independent given X2 and X3.

```
lmodel = lm(X2 ~ X3 + X1 + X4, data=data)
summary(lmodel)
```

```
##
## Call:
## lm(formula = X2 ~ X3 + X1 + X4, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90282 -0.15318  0.00188  0.15342  0.85952
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.001141   0.002247   0.508    0.612
## X3           0.012316   0.012177   1.011    0.312
## X1          -0.349303   0.006243 -55.950   <2e-16 ***
## X4          -0.352810   0.013891 -25.398   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2246 on 9996 degrees of freedom
## Multiple R-squared:  0.3841, Adjusted R-squared:  0.3839
## F-statistic:  2078 on 3 and 9996 DF,  p-value: < 2.2e-16
```

X3 is not significant while X1 and X4 are. This means X2 and X3 is independent given X1 and X4.

Fit with gRim

cannot install package, **remember to do it later**

```
glist <- list( 'X1', 'X2', 'X3', 'X4' )
ddd <- cov.wt(data, method="ML")
fit <- ggmfit(ddd$cov, ddd$n.obs, glist) # Estimate parameters using IPF
fit$K
```

```
##          X1       X2       X3       X4
## X1 5.513255  0.00000  0.00000  0.00000
## X2 0.000000 12.21077  0.00000  0.00000
## X3 0.000000  0.00000 20.54787  0.00000
## X4 0.000000  0.00000  0.00000 33.73434
```

It did not work. K has more elements equal to zero than the original one.

Problem 4

```
set.seed(123)
( Sig <- cbind(c(3,-1.4,0,0),c(-1.4,3,1.4,1.4),c(0,1.4,3,0),c(0,1.4,0,3)) )
```

```
##      [,1] [,2] [,3] [,4]
## [1,]  3.0 -1.4  0.0  0.0
## [2,] -1.4  3.0  1.4  1.4
## [3,]  0.0  1.4  3.0  0.0
## [4,]  0.0  1.4  0.0  3.0
```

```
data <- as.data.frame(mvrnorm(n=10000,mu=c(0,0,0,0),Sigma=Sig))
colnames(data) <- c("X1","X2","X3","X4")
```

a)

Correlation represented by graph

$X_1 \perp\!\!\!\perp X_3$ $X_1 \perp\!\!\!\perp X_4$ $X_2 \perp\!\!\!\perp X_4$ and they are not independent given X_2

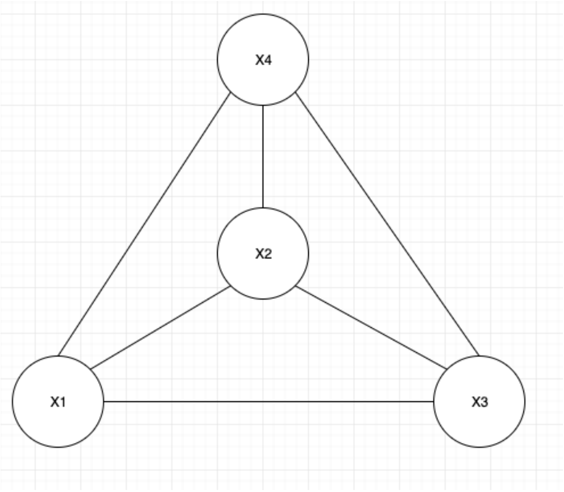
Correlation Matrix

```
solve(Sig)
```

```
##            [,1]       [,2]       [,3]       [,4]
## [1,]  0.5427350  0.4487179 -0.2094017 -0.2094017
## [2,]  0.4487179  0.9615385 -0.4487179 -0.4487179
## [3,] -0.2094017 -0.4487179  0.5427350  0.2094017
## [4,] -0.2094017 -0.4487179  0.2094017  0.5427350
```

b)

The moralized graph looks like



Every element of the precision matrix is not equal to 0 because every vertex is adjacent to another one.

It does not imply the correlation suggested in (a)

c)

```
glist <- list( 'X1', 'X2', 'X3', 'X4' )
ddd <- cov.wt(data, method="ML")
fit <- ggmfit(ddd$cov, ddd$n.obs, glist) # Estimate parameters using IPF
solve(fit$K)
```

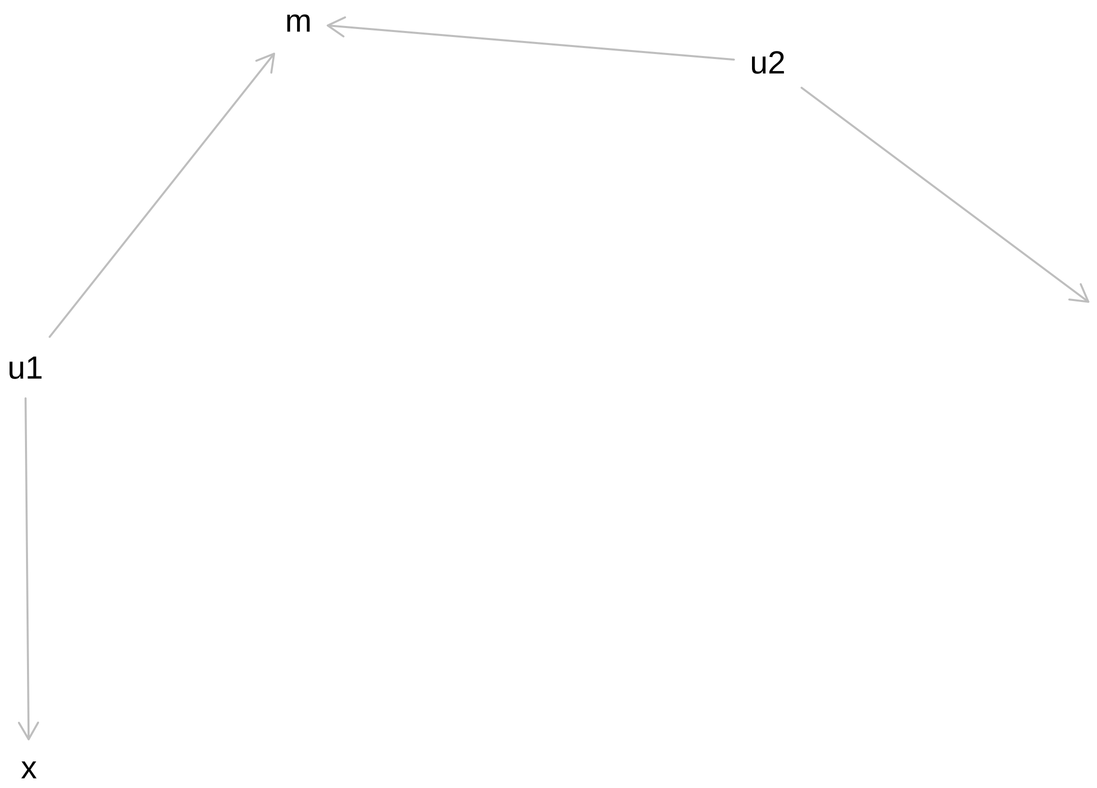
```
##          X1       X2       X3       X4
## X1 2.991722 0.000000 0.000000 0.000000
## X2 0.000000 2.959982 0.000000 0.000000
## X3 0.000000 0.000000 2.966525 0.000000
## X4 0.000000 0.000000 0.000000 3.077011
```

It is different from original covariance matrix as the elements on the diagonal are not the same.

Problem 5

```
g <- dagitty( "dag{ x <- u1; u1 -> m <- u2 ; u2 -> y }" )
df = simulateSEM(g, N = 1000, standardized = TRUE)
plot(g)
```

```
## Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to set you
```



```
reg = lm(y ~ x + m, data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x + m, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89042 -0.68302 -0.03076  0.67329  3.15642
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.02718    0.03113  -0.873   0.3829
## x           -0.06637    0.03043  -2.181   0.0294 *
## m            0.17789    0.03119   5.703 1.55e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9843 on 997 degrees of freedom
## Multiple R-squared:  0.03459,    Adjusted R-squared:  0.03265
## F-statistic: 17.86 on 2 and 997 DF,  p-value: 2.398e-08
```

```
confint(reg)
```

```
##                   2.5 %       97.5 %
## (Intercept) -0.0882695  0.033915940
## x           -0.1260776 -0.006660248
## m            0.1166794  0.239103260
```

The confidence interval of the effect (x) does not contain 0. This means the effect is negative.

Sufficient adjustment sets

```
adjustmentSets(g, exposure = 'x', outcome = 'y', type = 'all')
```

```
##  {}
## { u1 }
## { m, u1 }
## { u2 }
## { m, u2 }
## { u1, u2 }
## { m, u1, u2 }
```

One of the sufficient set is { m, u1, u2 }. The confidence interval of the effect (x) contains 0.

```
reg = lm(y ~ x + m + u1 + u2, data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x + m + u1 + u2, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6795 -0.6075 -0.0431  0.6084  3.1632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01797    0.02854  -0.630   0.5290
## x           -0.04940    0.02812  -1.757   0.0793 .
## m            0.05752    0.03121   1.843   0.0656 .
## u1          -0.05409    0.03018  -1.792   0.0734 .
## u2           0.42206    0.03109  13.577   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9019 on 995 degrees of freedom
## Multiple R-squared:  0.1912, Adjusted R-squared:  0.1879
## F-statistic:  58.8 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
confint(reg)
```

```
##                    2.5 %      97.5 %
## (Intercept) -0.073976553 0.038028095
## x           -0.104580175 0.005786093
## m           -0.003724192 0.118756250
## u1          -0.113306761 0.005132062
## u2           0.361056182 0.483060050
```

Another one of the sufficient set is { u2 }. The confidence interval of the effect (x) contains 0.

```
reg = lm(y ~ x  + u2, data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x + u2, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6052 -0.5939 -0.0578  0.5936  3.0800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01639    0.02858  -0.574   0.5664
## x           -0.05286    0.02786  -1.897   0.0581 .
## u2           0.44091    0.02937  15.012   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9033 on 997 degrees of freedom
## Multiple R-squared:  0.1869, Adjusted R-squared:  0.1853
## F-statistic: 114.6 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
confint(reg)
```

```
##                  2.5 %      97.5 %
## (Intercept) -0.0724672 0.039688379
## x           -0.1075186 0.001807503
## u2           0.3832788 0.498547229
```

The conclusion is that if the features input is a sufficient adjustment set plus exposure, the effect will not be significant.

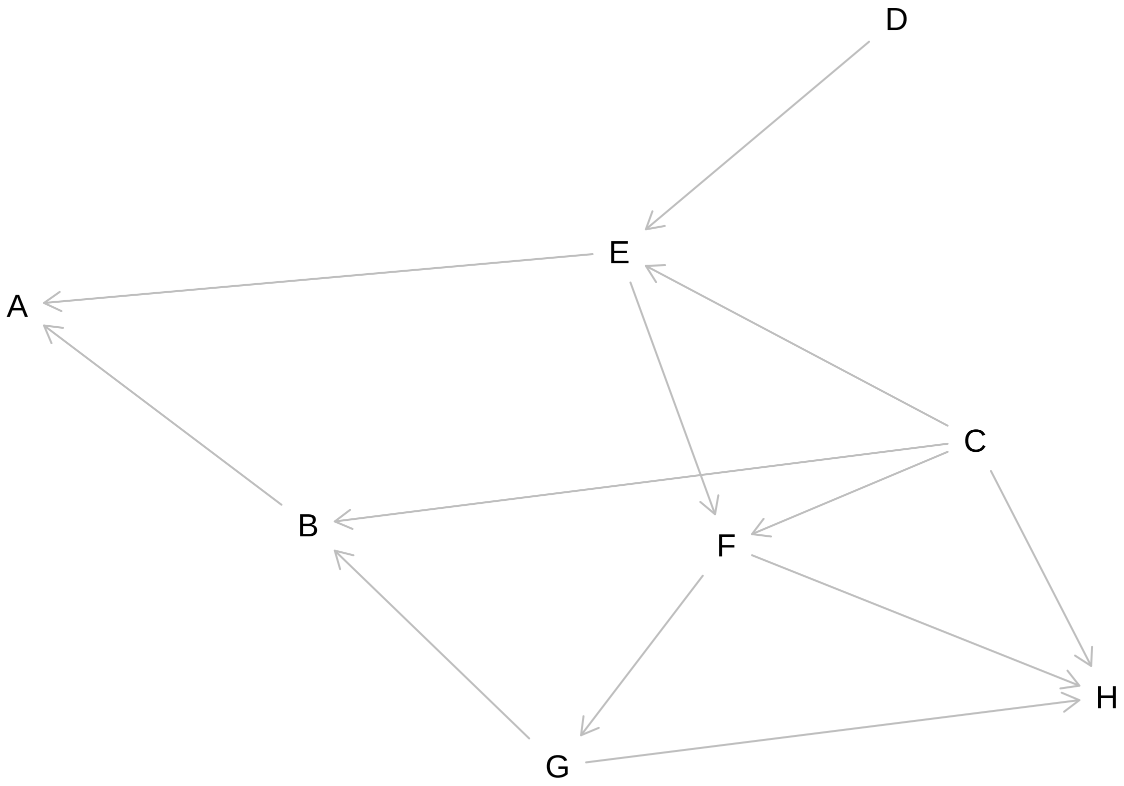
Problem 6

Construct the graph

```
g <- dagitty( "dag{
  D -> E -> A <- B <- G <- F -> H;
  G -> H;
  C -> H; C-> B; C->F; C-> E
  E -> F
}" )

df = simulateSEM(g, N = 10000, standardized = TRUE)
plot(g)
```

```
## Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to set you
```



Effects E on F

Sufficient adjustment sets

```
adjustmentSets(g, exposure = 'E', outcome = 'F', type = 'all')
```

```
## { C }
## { C, D }
```

Adjustment sets { C }

```
reg = lm(F ~ E + C, data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = F ~ E + C, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4891 -0.6332  0.0049  0.6412  3.7133
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.009042   0.009435  -0.958    0.338    
## E            0.273661   0.009488  28.844   <2e-16 ***
## C            0.207902   0.009518  21.842   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9434 on 9997 degrees of freedom
## Multiple R-squared:  0.1054, Adjusted R-squared:  0.1052 
## F-statistic: 589.1 on 2 and 9997 DF,  p-value: < 2.2e-16
```

Adjustment sets { C, D }

```
reg = lm(F ~ E + C + D, data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = F ~ E + C + D, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4676 -0.6338  0.0067  0.6427  3.7105 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -0.008974   0.009435  -0.951    0.342    
## E            0.282678   0.011696  24.169   <2e-16 ***
## C            0.208975   0.009553  21.876   <2e-16 ***
## D            0.015329   0.011628   1.318    0.187    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9434 on 9996 degrees of freedom
## Multiple R-squared:  0.1056, Adjusted R-squared:  0.1053 
## F-statistic: 393.3 on 3 and 9996 DF,  p-value: < 2.2e-16
```

All other variables

```
reg = lm(F ~ ., data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = F ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4704 -0.6250 -0.0003  0.6338  3.8190 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.009162   0.009385  -0.976    0.329    
## A            0.001863   0.010210   0.182    0.855    
## B            0.015339   0.009687   1.583    0.113    
## C            0.234798   0.009844  23.851   <2e-16 ***
## D            0.011964   0.011568   1.034    0.301    
## E            0.276885   0.012342  22.434   <2e-16 ***
## G            0.006011   0.009948   0.604    0.546    
## H           -0.101338   0.010069 -10.065   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.9382 on 9992 degrees of freedom
## Multiple R-squared:  0.1158, Adjusted R-squared:  0.1152 
## F-statistic:   187 on 7 and 9992 DF,  p-value: < 2.2e-16
```

Variance of estimates All other variables < { C } < { C, D }

Effects B on A

Sufficient adjustment sets

```
adjustmentSets(g, exposure = 'B', outcome = 'A', type = 'all')
```

```
## { E }
## { C, E }
## { D, E }
## { C, D, E }
## { C, F }
## { C, D, F }
## { E, F }
## { C, E, F }
## { D, E, F }
## { C, D, E, F }
## { C, G }
## { C, D, G }
## { E, G }
## { C, E, G }
## { D, E, G }
## { C, D, E, G }
## { C, F, G }
## { C, D, F, G }
## { E, F, G }
## { C, E, F, G }
## { D, E, F, G }
## { C, D, E, F, G }
## { E, H }
## { C, E, H }
## { D, E, H }
## { C, D, E, H }
## { C, F, H }
## { C, D, F, H }
## { E, F, H }
## { C, E, F, H }
## { D, E, F, H }
## { C, D, E, F, H }
## { C, G, H }
## { C, D, G, H }
## { E, G, H }
## { C, E, G, H }
## { D, E, G, H }
## { C, D, E, G, H }
## { C, F, G, H }
## { C, D, F, G, H }
## { E, F, G, H }
## { C, E, F, G, H }
## { D, E, F, G, H }
## { C, D, E, F, G, H }
```

Sufficient adjustment set { E }

```
reg = lm(A ~ B + E, data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = A ~ B + E, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0834 -0.6134 -0.0121  0.6227  3.5832
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003610   0.009195  -0.393    0.695
## B            0.064007   0.009255   6.916 4.93e-12 ***
## E            0.405688   0.009182  44.181  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9193 on 9997 degrees of freedom
## Multiple R-squared:  0.1672, Adjusted R-squared:  0.167
## F-statistic:  1004 on 2 and 9997 DF,  p-value: < 2.2e-16
```

Sufficient adjustment set { C, D, E, G, H }

```
reg = lm(A ~ B + C + D + E + F + G , data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = A ~ B + C + D + E + F + G, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0833 -0.6107 -0.0132  0.6209  3.5697
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0036327  0.0091964  -0.395    0.693
## B            0.0611686  0.0094741   6.456 1.12e-10 ***
## C           -0.0147689  0.0095744  -1.543    0.123
## D           -0.0052588  0.0113331  -0.464    0.643
## E            0.4008581  0.0117258  34.186  < 2e-16 ***
## F            0.0004585  0.0097537   0.047    0.963
## G           -0.0076699  0.0093327  -0.822    0.411
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9193 on 9993 degrees of freedom
## Multiple R-squared:  0.1675, Adjusted R-squared:  0.167
## F-statistic:   335 on 6 and 9993 DF,  p-value: < 2.2e-16
```

All other variables

```
reg = lm(A ~ ., data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = A ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0708 -0.6121 -0.0147  0.6206  3.5601
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003560   0.009196  -0.387   0.6987
## B            0.061155   0.009474   6.455 1.13e-10 ***
## C           -0.018264   0.009915  -1.842   0.0655 .
## D           -0.004895   0.011336  -0.432   0.6659
## E            0.401103   0.011727  34.204  < 2e-16 ***
## F            0.001789   0.009803   0.182   0.8552
## G           -0.003855   0.009748  -0.396   0.6925
## H            0.013435   0.009915   1.355   0.1754
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9193 on 9992 degrees of freedom
## Multiple R-squared:  0.1676, Adjusted R-squared:  0.167
## F-statistic: 287.5 on 7 and 9992 DF,  p-value: < 2.2e-16
```

Variance of estimates All other variables > { C, D, E, F, G } > {E}.

Two results are on the contrary. The explanation is that B has only one path directly out to A while E has one to F and another one to A.