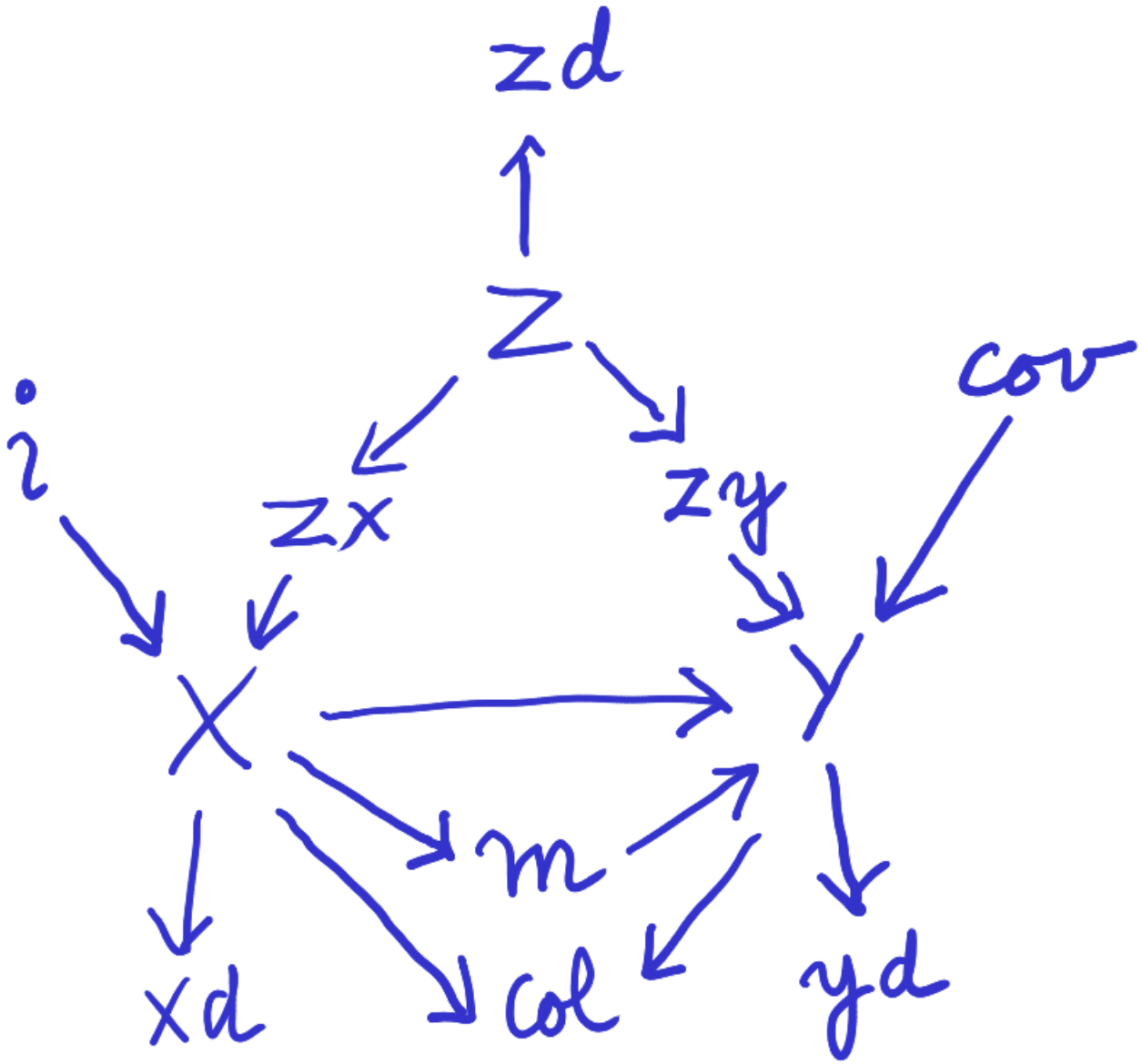
IV or Regression

2022-03-25

Contents

1 The Causal Zoo




```
{
  nams <- c('z','zx','zy','cov','x','y',
            'm','i', 'xd', 'yd', 'zd',
            'col', 'i2')
  mat <- matrix(0, length(nams), length(nams))
  rownames(mat) <- nams
  colnames(mat) <- nams

  # confounding back-door path

  mat['zx','z'] <- 3
  mat['zy','z'] <- 3
  mat['x','zx'] <- 1
```

```
mat['y','zy'] <- 2

# direct effect of x on y

mat['y','x'] <- 3

# indirect effect: Note that the causal effect is 3 + 1x1 = 4

mat['m','x'] <- 1
mat['y','m'] <- 1

# Instrumental variable

mat['x','i'] <- 1

# Secondary instrument

mat['zx','i2'] <- 1

# 'Covariate'

mat['y','cov'] <- 2

# descendant of X

mat['xd','x'] <- 1

# descendant of Y

mat['yd','y'] <- 1

# descendant of z -- imperfect control

mat['zd','z'] <- 2

# collider

mat['col','y'] <- 1
mat['col','x'] <- 1


# independent SD of error for every variable

diag(mat) <- 1

# but make SD of Y smaller

mat['y','y'] <- .01
mat['i2','i2'] <- 3
mat['i','i'] <- 2


mat # not in lower diagonal form
```

```
  dag <- permute_to_dag(mat) # can be permuted to lower-diagonal form
  dag # this allows us to iteratively work out the covariance matrix
}
```

```
     i2 i z zx x m cov zy    y col zd yd xd
i2    3 0 0  0 0 0   0  0 0.00   0  0  0  0
i     0 2 0  0 0 0   0  0 0.00   0  0  0  0
z     0 0 1  0 0 0   0  0 0.00   0  0  0  0
zx    1 0 3  1 0 0   0  0 0.00   0  0  0  0
x     0 1 0  1 1 0   0  0 0.00   0  0  0  0
m     0 0 0  0 1 1   0  0 0.00   0  0  0  0
cov   0 0 0  0 0 0   1  0 0.00   0  0  0  0
zy    0 0 3  0 0 0   0  1 0.00   0  0  0  0
y     0 0 0  0 3 1   2  2 0.01   0  0  0  0
col   0 0 0  0 1 0   0  0 1.00   1  0  0  0
zd    0 0 2  0 0 0   0  0 0.00   0  1  0  0
yd    0 0 0  0 0 0   0  0 1.00   0  0  1  0
xd    0 0 0  0 1 0   0  0 0.00   0  0  0  1
attr(,"class")
[1] "dag"    "matrix" "array"
```

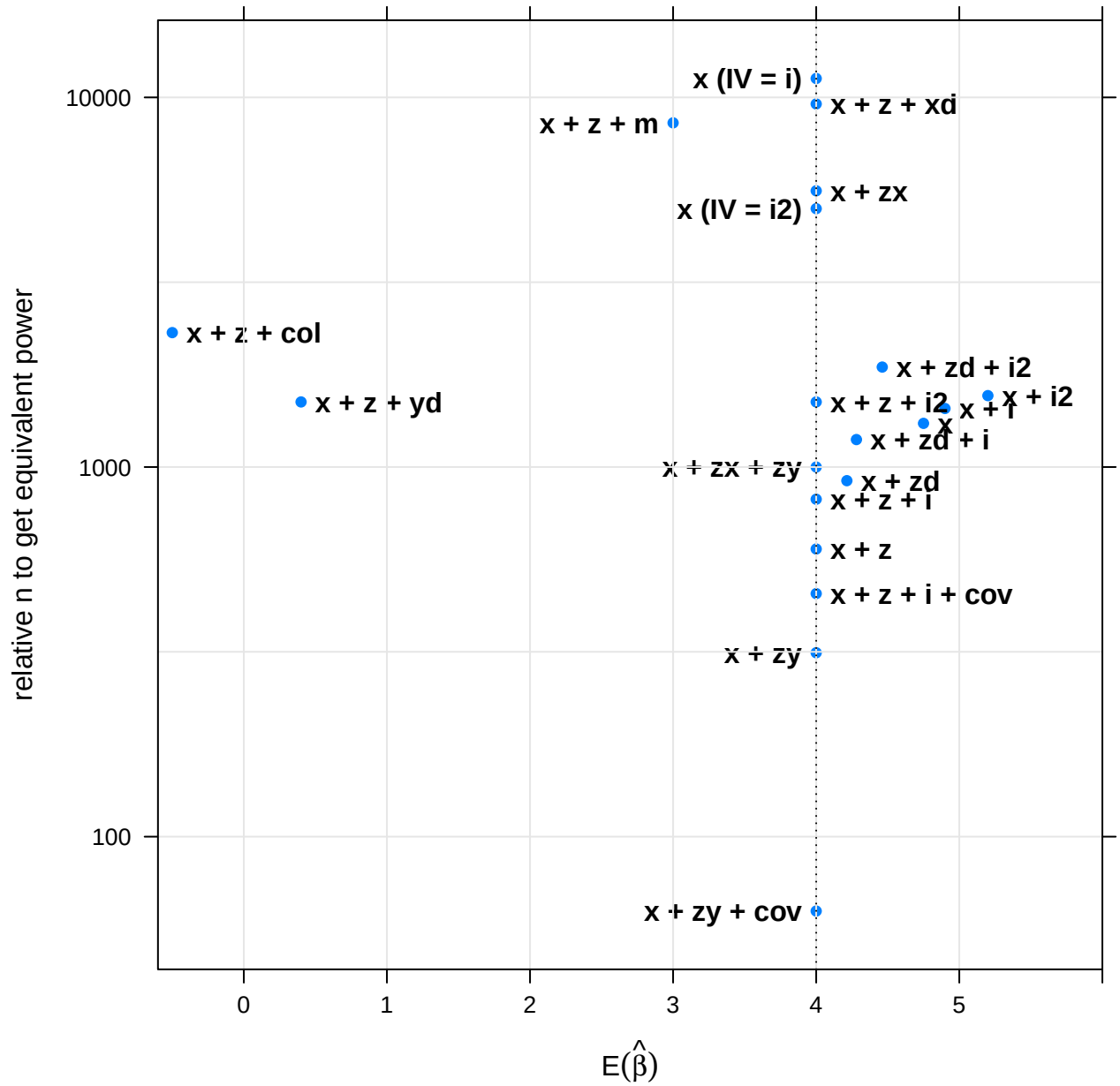
Some models to try:

```
fmlas <- list(
  y ~ x,              # with confounding
  y ~ x + z,          # unconfounded
  y ~ x + zy,         # unconfounded using generating model
  y ~ x + zx,         # unconfounded using assignment model
  y ~ x + zx + zy,    # 'doubly robust'
  y ~ x + zy + cov,   # adding a covariate unrelated to x
  y ~ x + z + m,     # adding a mediator
  y ~ x + z + xd,    # adding a descendant of X
  y ~ x + z + yd,    # adding a descendant of Y
  y ~ x + z + col,   # adding a collider
  y ~ x + z + i,     # adding an instrumental variable
  y ~ x + z + i2,     # adding a secondary instrumental variable
  y ~ x + z + i + cov,  # adding an instrumental variable and a covariate
  y ~ x + i,          # using an instrumental variable as a control
  y ~ x + i2,          # using an instrumental variable as a control
  y ~ x + zd,         # imperfect control for confounding
  y ~ x + zd + i,      # bias amplification
  y ~ x + zd + i2,     # bias amplification
  y ~ x | i,           # instrumental variable
  y ~ x | i2          # secondary instrumental variable
)

fmlas %>%
  lapply(coefx, dag) %>%
  lapply(as.data.frame) %>%
  do.call(rbind.data.frame, .) -> df

pdf <- df
sapply(pdf, is.numeric) %>%
  {pdf[,.] <<- round(pdf[,.], 2)}
pdf[, c(5,1,4,2,3)] %>% print(row.names=F)
```

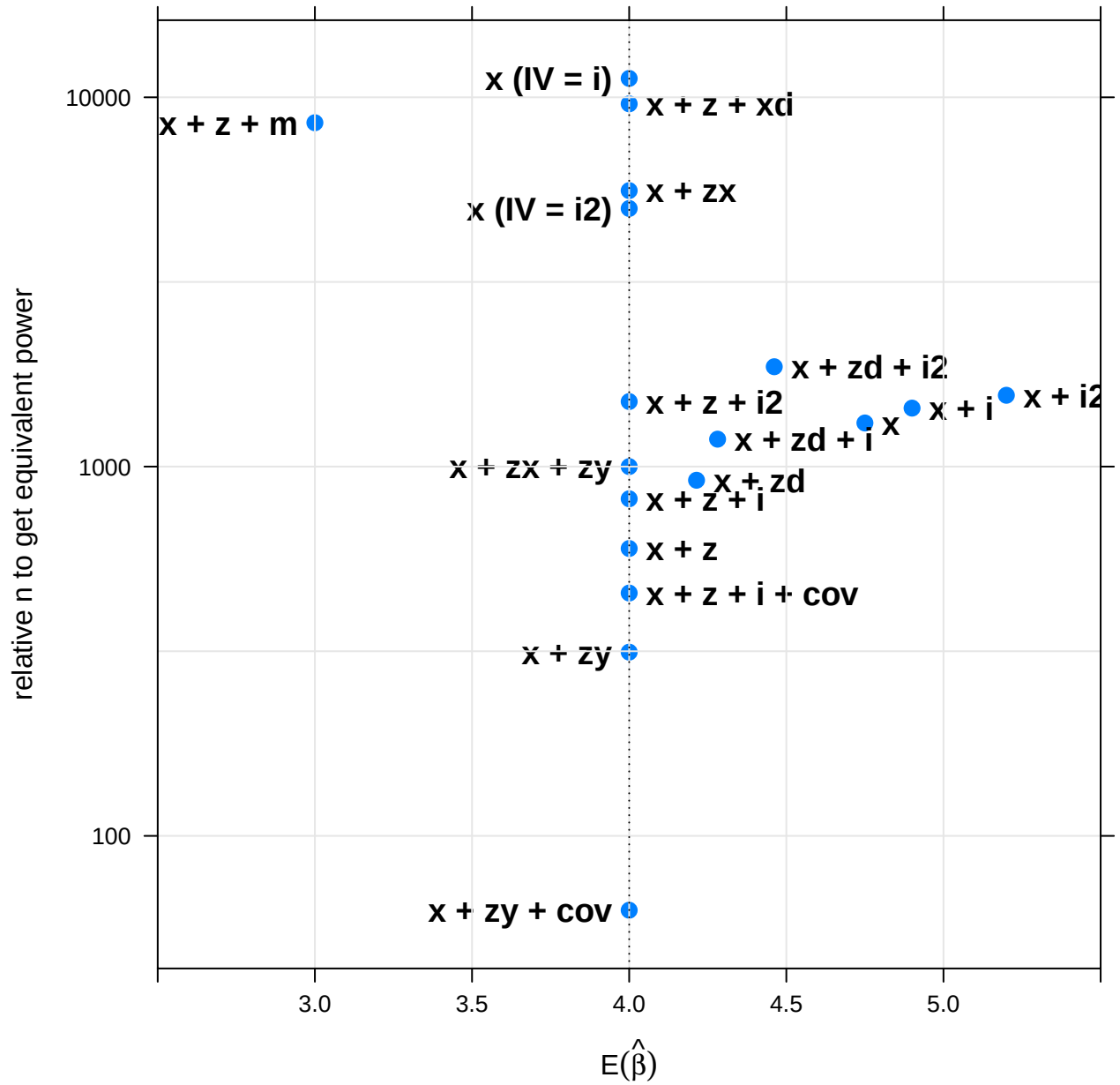
```
                    label beta_x sd_factor sd_e sd_x_avp
                    y ~ x   4.75      1.15 5.61     4.90
                y ~ x + z   4.00      0.77 3.00     3.87
               y ~ x + zy   4.00      0.56 2.24     3.99
               y ~ x + zx   4.00      2.36 5.29     2.24
           y ~ x + zx + zy   4.00      1.00 2.24     2.24
          y ~ x + zy + cov   4.00      0.25 1.00     3.99
             y ~ x + z + m   3.00      2.92 2.83     0.97
            y ~ x + z + xd   4.00      3.10 3.00     0.97
            y ~ x + z + yd   0.40      1.22 0.95     0.77
           y ~ x + z + col  -0.50      1.52 0.95     0.62
             y ~ x + z + i   4.00      0.90 3.00     3.32
            y ~ x + z + i2   4.00      1.22 3.00     2.45
      y ~ x + z + i + cov   4.00      0.67 2.24     3.32
               y ~ x + i   4.90      1.20 5.37     4.47
              y ~ x + i2   5.20      1.25 4.84     3.87
              y ~ x + zd   4.21      0.96 3.93     4.10
            y ~ x + zd + i   4.28      1.09 3.90     3.58
           y ~ x + zd + i2   4.46      1.37 3.81     2.79
            y ~ x (IV = i)   4.00      3.35 6.71     2.00
           y ~ x (IV = i2)   4.00      2.24 6.71     3.00
```

```r
df <- within(
  df,
  {
    pos <- ifelse(grepl('IV|zy|m',label), 2, 4)
  }
)
```

```
xyplot(log10(sd_factor^2) ~ beta_x, df , font = 2,
       # scales = list(y = list(log = 10)),
       scales = list(y = list(at=seq(-3,3), labels=10^(3+seq(-3,3)))),
       xlim = c(-.6, 6),
       pch = 16,
       xlab = TeX('$E(\\hat{\\beta})$'),
       ylab = TeX('relative $n$ to get equivalent power'),
       labs = sub('y ~ ','',df$label),
       pos = df$pos) +
  layer(panel.text(..., labels = labs, pos = pos)) +
  layer(panel.grid(h=-1,v=-1)) +
  layer(panel.abline(v = 4, lty = 3))
```



```
xyplot(log10(sd_factor^2) ~ beta_x, df , font = 2, cex = 1.2,
       # scales = list(y = list(log = 10)),
       scales = list(y = list(at=seq(-3,3), labels=10^(3+seq(-3,3)))),
       xlim = c(2.5, 5.5),
       pch = 16,
       xlab = TeX('$E(\\hat{\\beta})$'),
       ylab = TeX('relative $n$ to get equivalent power'),
       labs = sub('y ~ ','',df$label),
       pos = df$pos) +
  layer(panel.text(..., labels = labs, pos = pos)) +
  layer(panel.grid(h=-1,v=-1)) +
  layer(panel.abline(v = 4, lty = 3))
```



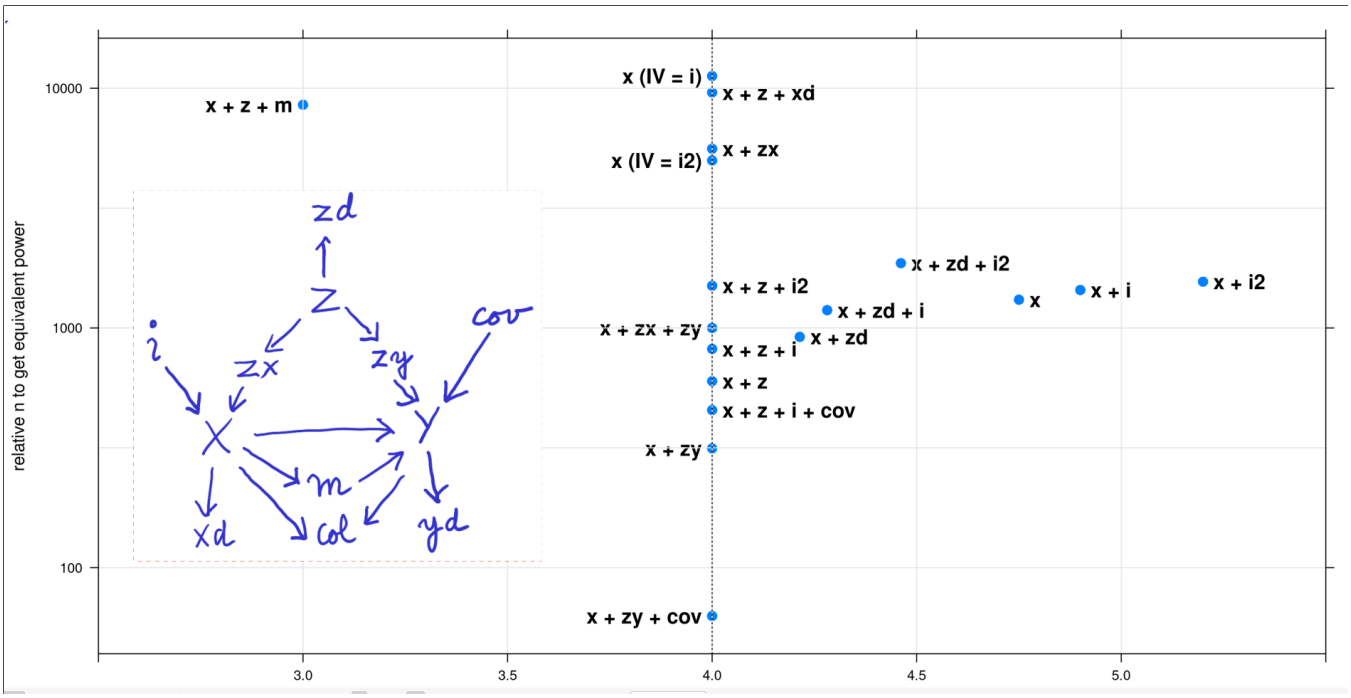
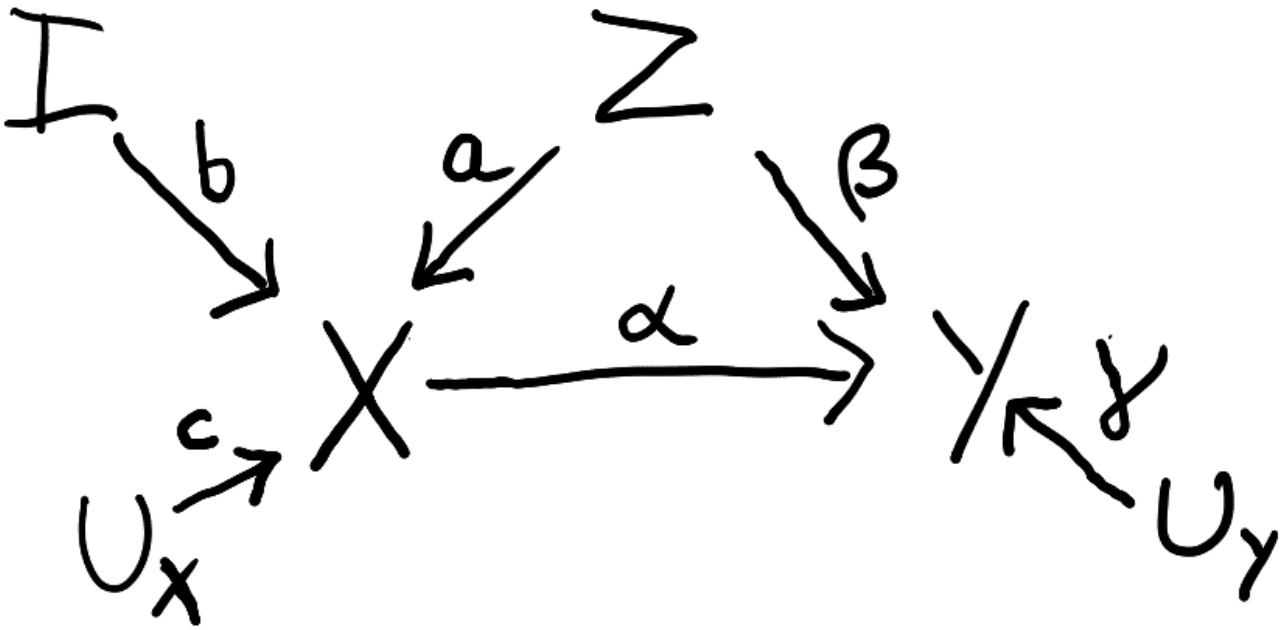


Figure 1: width='50%'

2 What's Happening?



Let's assume multivariate normality and build a variance matrix for Z, I, X, Y.

We can scale I, Z and X so they have unit variance and zero means. This eliminates irrelevant nuisance parameters.

Since I is an instrument for the confounding effect of Z:

$$Var\begin{pmatrix} Z \\ I \\ X \end{pmatrix} = \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & b \\ a & b & 1 \end{pmatrix}$$

with $a^2 + b^2 \leq 1$.

Focus first on the *assignment model*, i.e. the model that determines the value of X from the values of Z, I and U_X.

Letting $c^2 = 1 - a^2 - b^2$, c^2 represent the portion of the variance in X that is not attributed to the instrument, I, nor to the confounder, Z, define

$$\rho_I = \frac{b^2}{b^2 + c^2}$$

the proportion of the variance in X not due to Z that is 'explained' by I.

For an instrument that captures all of the variation not due to the confounder, $c^2 = 0$ and $\rho_I = 1$.

Focusing next on the model generating Y, let

$$Y = \alpha X + \beta Z + \gamma\varepsilon$$

with $\varepsilon \sim \mathrm{N}(0, 1)$, independent of other variables.

The variance matrix is:

$$Var\begin{pmatrix} Z \\ I \\ X \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 & a & a\alpha + \beta \\ 0 & 1 & b & b\alpha \\ a & b & 1 & \alpha + a\beta \\ a\alpha + \beta & b\alpha & \alpha + a\beta & v_{yy} \end{pmatrix}$$

where $v_{yy} = \alpha^2 + \beta^2 + 2a\alpha\beta + \sigma_\varepsilon^2$

We can verify that the regression coefficients for the regression of Y on X and Z are

$$\begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}^{-1} \begin{pmatrix} \alpha + a\beta \\ a\alpha + \beta \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

The variance of the least-squares estimator of α based on a regression on X and the confounder Z is:

$$\begin{aligned} \text{Var}(\hat{\alpha}) &\approx \frac{1}{n} \frac{\sigma_\epsilon^2}{1 - a^2} \\ &= \frac{1}{n} \frac{\gamma^2}{b^2 + c^2} \end{aligned}$$

The asymptotic expectation of the instrumental variable estimator $\tilde{\alpha}$ is

$$\sigma_{IX}^{-1} \sigma_{IY} = \frac{1}{b} \times b\alpha = \alpha$$

The variance of $\tilde{\alpha}$ is (Fox 2016, 241):

$$\text{Var}(\tilde{\alpha}) \approx \frac{1}{n} \sigma_{\epsilon IV}^2 \sigma_{IX}^{-1} \sigma_{II} \sigma_{XI}^{-1} \quad = \frac{1}{n} (\beta^2 + \gamma^2) \frac{1}{b^2}$$

Thus, the variance inflation factor – which is the same as the 'sample size inflation factor to achieve the same power' – using IV estimation instead of controlling for a confounder (assuming that both approaches are available) is:

$$\begin{aligned} IVVIF &= \frac{\text{Var}(\tilde{\alpha})}{\text{Var}(\hat{\alpha})} \\ &= \frac{\beta^2 + \gamma^2}{b^2} \Big/ \frac{\gamma^2}{b^2 + c^2} \\ &= \frac{\beta^2 + \gamma^2}{\gamma^2} \Big/ \frac{b^2}{b^2 + c^2} \\ &= 1 \Big/ \left(\frac{\gamma^2}{\gamma^2 + \beta^2} \times \frac{b^2}{b^2 + c^2} \right) \\ &= \left(1 + \frac{\beta^2}{\gamma^2} \right) \times \left(1 + \frac{c^2}{b^2} \right) \\ &= \frac{1}{1 - R^2_{Y,Z|X}} \times \frac{1}{R^2_{X,I|Z}} \end{aligned}$$

The first term is structural in the sense that it is a consequence of the problem, specifically the degree of confounding relative to the residual error variance in the model. For a given problem, the IV has no impact on this, so it represents a lower bound for the IVVIF. The second term clarifies that it is not the *correlation of the IV with X* directly that affects the IVVIF, but its **partial correlation** adjusted for the relationship of X with confounders.

References

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models.* 3rd ed. Sage Publications.