EPSY 8266 Lab 2: Simulation in R

Chris Desjardins
2/4/2019

Purpose

The purpose of this lab is get you familar with using simulation in R. Specifically, we will use simulation to understand when we need to control for a variable W in order to get an unbiased total estimate of X on Y. We will use `lavaan` to do this, but you could just as easily do this just using regression and `lm`.

Lab Data

For this lab, you'll be generating the random data through Monte Carlo simulation.

Lab Format

For the simulation questions, I'll give you most of the code you'll need, as well as the parameters, and you'll tell me whether or not we need to include W in the model if we care just about getting an unbiased estimate of the total effect of (the sum of the direct and indirect effects) X on Y.

The total effect of X on Y is defined as:

Total Effect = Direct Effect + Indirect Effect

- The direct effect is just the partial regression coefficient, often referred to as **c**, and in Model 1 will be equal to .5 (see Figure 1 below).
- The indirect effect is the product of the effect of X on W (**a**) multiplied by the effect of W on Y (**b**).

Why do these equal to the total effect of X on Y? Imagine the following two linear regressions:

$$\hat{W} = aX \tag{1}$$

$$\hat{Y} = bW + cX \tag{2}$$

We can interpret **a** in Equation 1 as a one-unit increase in X results in an **a** increase in W. We can then substitute **a** into Equation 2 for W and we can see that it becomes **a***b, which is how much Y goes for a one-unit increase in X through W.

Therefore, a one-unit increase in X results in a direct effect of **c** on Y and an indirect effect of **a***b on Y through W.

Looking at Model 1 in Figure 1, we see that the effect of X on W is .4 an the effect of W on Y is .3.

Question 1: Calculate the indirect effect and the total effect of X on Y for Model 1.

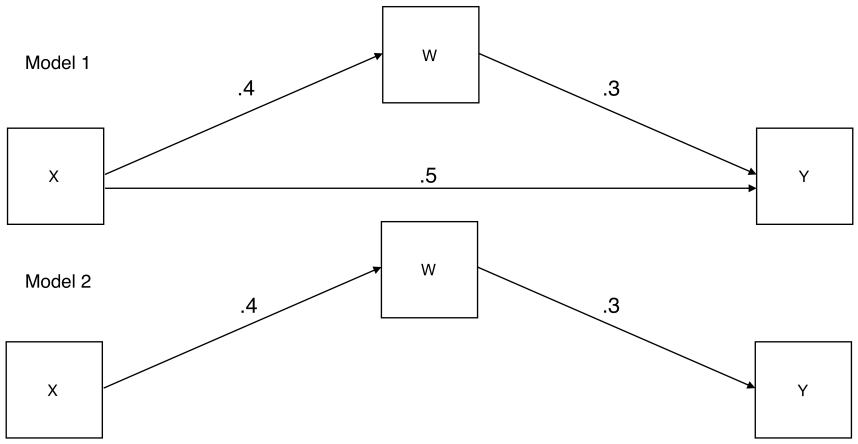

Figure 1: Models 1 (top) and 2 (bottom).

For Model 1, the true model is that X causes W and X and W bpth cause Y (note the direction of the paths). In contrast for Model 2, the true model is that X causes W and W causes Y. Note that there is no direct path from X to Y, however, does this mean that X doesn't affect Y?

Let's begin by generating some data from Model 1, fitting a regression, omitting W from our model, and seeing if we can recover our answer for the total effect that we provided for Question 1.

```
# load lavaan
library(lavaan)

## This is lavaan 0.6-3
## lavaan is BETA software! Please report any bugs.
```

```
set.seed(123512)
#
#
# Setting up conditions for model 1
#
#

# relationship between X and W
a <- .4

# relationship between W and Y
b <- .3

# relationship between X and Y
c <- .5

# how many observations should we generate.
# the more we generate the closer we will get to the true parameters
```

```r
# (i.e, the small error we'll have) recall sampling distributions.
n <- 500

# let's generate X to be a random normal variable
X <- rnorm(n = n, mean = 0, sd = 1)

# let's generate W
W <- a*X + rnorm(n, mean = 0, sd = sqrt(1 - a^2))
# - The .4 is the standardized regression weight between X and W
# - The rnorm() stuff adds the residual variance to make the correlation between
# - X and W .4
# - You can verify this by setting n to a really large number and doing
# - cor(X, W)

# Now let's generate Y
Y <- b*W + c*X + rnorm(n, mean = 0, sd = sqrt(1 - (b^2 + c^2 + 2*b*c*a)))
dat <- data.frame(X, W, Y)

# And let's fit a path model and omit W
mod <- '
Y ~ X
'
fit <- sem(model = mod, data = dat)
summary(fit)
params <- parameterEstimates(fit)
params[params$lhs == "Y" & params$rhs == "X", "est"]
```

Question 2: What was the estimated effect? How close was this value to what you calculated in Question 1? Would you call this a small or a big difference?

In order to understand if there is bias from omitting W on the total effect of X on Y, we need to repeat this process multiple times. This is what's known as a simulation. We are doing this by generating random data from our model. This is what makes it Monte Carlo.

Let's write a function that will repeat this over and over again and return the total effect, so we don't have to run that above code 2000 times!

```r
set.seed(125312)

# how many replicates should we use?
nsim <- 2000

# run the simulation
# notice that everything in the expr argument was defined earlier.
runSim <- replicate(nsim, expr = {

  # all these lines are the same as above
  X <- rnorm(n = n, mean = 0, sd = 1)
  W <- a*X + rnorm(n, mean = 0, sd = sqrt(1 - a^2))
  Y <- b*W + c*X + rnorm(n, mean = 0, sd = sqrt(1 - (b^2 + c^2 + 2*b*c*a)))
  dat <- data.frame(X, W, Y)
  fit <- sem(model = mod, data = dat)
```

```
  params <- parameterEstimates(fit)
  params[params$lhs == "Y" & params$rhs == "X", "est"]
  })

# let's plot the results
hist(runSim)

# Now calculate the mean
mean(runSim)
```

Question 3: Describe the distribution of *runSim*.

Question 4: How big is the difference between the mean of our simulation and what you calculated in Question 1? Would you say this is big or small? Would you say that the total effect is unbiased when W is omitted?

Question 5: What is the total effect of X on Y in Model 2 (Figure 1)? What is the indirect effect and what is the direct effect?

Now, let's run Model 2. We need to change either a, b, or c below. Please change the correct one in the code below.

```
set.seed(125312)
#
#
# Setting up conditions for model 2
#
#

#
# Change one of these parameters
#
# relationship between X and w
a <- .4

# relationship between W and Y
b <- .3

# relationship between X and Y
c <- .5
#
#
#

# how many observations should we generate.
# the more we generate the closer we will get to the true parameters
# (i.e, the small error we'll have) recall sampling distributions.
n <- 500
# how many replicates should we use?
nsim <- 2000

# run the simulation
# notice that everything in the expr argument was defined earlier.
```

```
runSim <- replicate(nsim, expr = {

  # all these lines are the same as above
  X <- rnorm(n = n, mean = 0, sd = 1)
  W <- a*X + rnorm(n, mean = 0, sd = sqrt(1 - a^2))
  Y <- b*W + c*X + rnorm(n, mean = 0, sd = sqrt(1 - (b^2 + c^2 + 2*b*c*a)))
  dat <- data.frame(X, W, Y)
  fit <- sem(model = mod, data = dat)
  params <- parameterEstimates(fit)
  params[params$lhs == "Y" & params$rhs == "X", "est"]
  })

# let's plot the results
hist(runSim)

# Now calculate the mean
mean(runSim)
```

Question 6: What is the mean from the simulation? Do we need W in the model to obtain an unbiased estimate of the total effect of X on Y for this model?

Question 7: For Model 3, in Figure 2, what is the total effect of X on Y? What is the direct effect? What is the indirect effect?

Please modify the parameters below for Model 3 (if necessary).

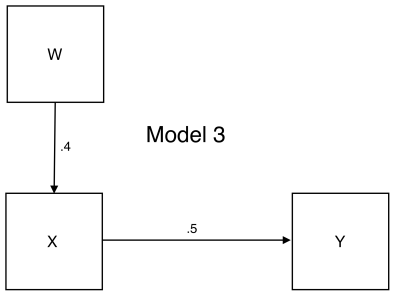

Figure 2: Model 3.

```
set.seed(125312)
#
#
# Setting up conditions for model 3
#
#

#
# Change one or more of these (if necessary)
#
# relationship between W and X
a <- .4
```

```r
# relationship between W and Y
b <- .3

# relationship between X and Y
c <- .5
#
#
#

# how many observations should we generate.
# the more we generate the closer we will get to the true parameters
# (i.e, the small error we'll have) recall sampling distributions.
n <- 500

# how many replicates should we use?
nsim <- 2000

# run the simulation
runSim <- replicate(nsim, expr = {

  # all these lines are the same as above
  W <- rnorm(n = n, mean = 0, sd = 1)
  X <- a*W + rnorm(n, mean = 0, sd = sqrt(1 - a^2))
  Y <- b*W + c*X + rnorm(n, mean = 0, sd = sqrt(1 - (b^2 + c^2 + 2*b*c*a)))
  dat <- data.frame(X, W, Y)
  fit <- sem(model = mod, data = dat)
  params <- parameterEstimates(fit)
  params[params$lhs == "Y" & params$rhs == "X", "est"]
  })

# let's plot the results
hist(runSim)

# Now calculate the mean
mean(runSim)
```

Question 8: What is the mean from the simulation? Do we need W in the model to obtain an unbiased estimate of the total effect of X on Y?

Question 9: For Model 4, in Figure 3, what is the total effect of X on Y? What is the direct effect? What is the indirect effect?

Please modify the parameters below for Model 4 (if necessary).

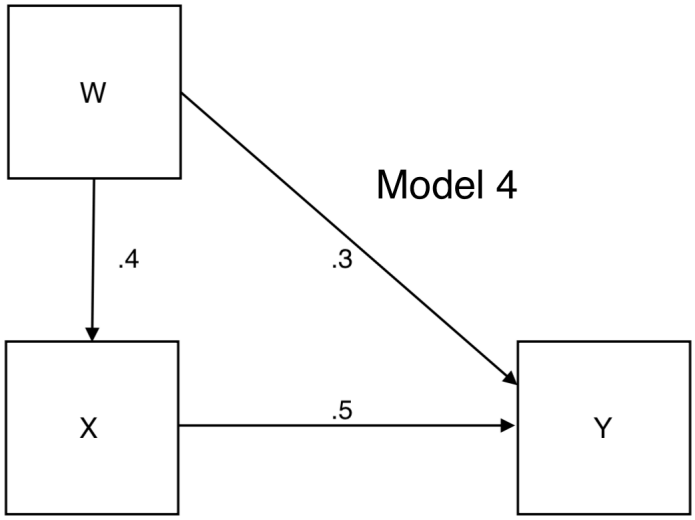

Figure 3: Model 4.

```r
set.seed(125312)
#
#
# Setting up conditions for model 4
#
#


#
# Change one or more of these  (if necessary)
#
# relationship between W and X
a <- .4

# relationship between W and Y
b <- .3

# relationship between X and Y
c <- .5
#
#
#


# how many observations should we generate.
# the more we generate the closer we will get to the true parameters
# (i.e, the small error we'll have) recall sampling distributions.
n <- 500

# how many replicates should we use?
nsim <- 2000

# run the simulation
# notice that everything in the expr argument was defined earlier.
```

```r
runSim <- replicate(nsim, expr = {

  # all these lines are the same as above
  W <- rnorm(n = n, mean = 0, sd = 1)
  X <- a*W + rnorm(n, mean = 0, sd = sqrt(1 - a^2))
  Y <- b*W + c*X + rnorm(n, mean = 0, sd = sqrt(1 - (b^2 + c^2 + 2*b*c*a)))
  dat <- data.frame(X, W, Y)
  fit <- sem(model = mod, data = dat)
  params <- parameterEstimates(fit)
  params[params$lhs == "Y" & params$rhs == "X", "est"]
  })

# let's plot the results
hist(runSim)

# Now calculate the mean
mean(runSim)
```

Question 10: What is the mean from the simulation? Do we need W in the model to obtain an unbiased estimate of the total effect of X on Y? HINT: What model does this value look like?

Question 11: For Model 5, in Figure 5, what is the total effect of X on Y? What is the direct effect? What is the indirect effect?

For the final model, Model 5, we will run the simulations two-ways. 1) with W in the model as a predictor and 2) without W in the model.

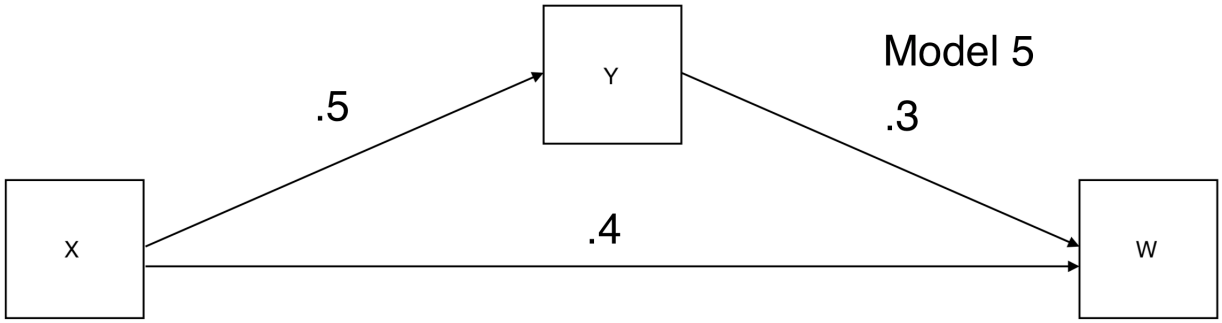

Figure 4: Model 5.

```r
set.seed(125312)
#
#
# Setting up conditions for model 5
#
#

# relationship between X and W
a <- .4
```

```r
# relationship between W and Y
b <- .3

# relationship between X and Y
c <- .5
#
#
#

# how many observations should we generate.
# the more we generate the closer we will get to the true parameters
# (i.e, the small error we'll have) recall sampling distributions.
n <- 500

set.seed(125312)

# how many replicates should we use?
nsim <- 2000

# create the misspecified model - this one includes W
mod.w <- '
  Y ~ X + W
'

# run the simulation
# notice that everything in the expr argument was defined earlier.
runSim <- replicate(nsim, expr = {

  # all these lines are the same as above
  X <- rnorm(n = n, mean = 0, sd = 1)
  Y <- c*X + rnorm(n, mean = 0, sd = sqrt(1 - c^2))
  W <- b*Y + a*X + rnorm(n, mean = 0, sd = sqrt(1 - (b^2 + a^2 + 2*b*c*a)))
  dat <- data.frame(X, W, Y)
  fit <- sem(model = mod, data = dat)
  fit.w <- sem(model = mod.w, data = dat)
  params <- parameterEstimates(fit)
  params.w <- parameterEstimates(fit.w)
  c(params[params$lhs == "Y" & params$rhs == "X", "est"],
    params.w[params.w$lhs == "Y" & params.w$rhs == "X", "est"])
  })

# Now print the means
rownames(runSim) <- c("Correct", "Misspecified")
rowMeans(runSim)
```

Question 12: What is the mean from the simulation for the Correct Model? Is it an unbiased estimate of X on Y? What is the mean from the simulation for the Misspecified Model? Is it an unbiased estimate of X on Y?

Question 13: For which model(s) is it fine to omit W? For which model(s) must we include W? For which model(s) must we make sure not to include W? This question doesn't require any additional analysese, just summarizing your findings.

These models highlight the importance of carefully considering the relationships between 3 variables! At a later date, I will share slides on the course website about other relationships.