1 Inducing correlation via a shared random variable

I will do this for Normals, as it is one of the few cases where you can see explicitly what is going on via formulas.

1.1 Preliminaries

For two random variables A and B we have the definitions:

$$Cov(A, B) = E(A\,B) - E(A)E(B).$$

$$Cor(A, B) = \frac{Cov(A, B)}{\sqrt{Var(A)Var(B)}}.$$

If two random variables A and B are independent then

$$E(A\,B) = E(A)E(B).$$

1.2 Set up the analysis

We introduce three random variables, U, V and T which are defined to be independent of one another.

For simplicity let U be normal, mean 0 and variance 1, that is $U \sim N(0, 1)$.

Let V be normal, mean 0 and variance 1, $V \sim N(0, 1)$.

We now introduce the shared variable T which is Normal mean 0, variance σ_T^2.

Define X = U + T and Y = V + T. Note that these two new random variables X and Y share the common variable T.

1.3 What's the correlation between X and Y?

First work out the covariance:

$$\begin{aligned} Cov(X, Y) &= Cov(U + T, V + T) && (1)\\ &= E((U + T)(V + T)) - E(U + T)E(V + T) && (2)\\ &= E((U + T)(V + T)) - 0 && (3) \end{aligned}$$

because all of U, V and T are mean 0.

Now

$$\begin{aligned} E((U + T)(V + T)) &= E(UV) + E(TV) + E(TV) + E(T^2) && (4)\\ &= 0 + 0 + 0 + E(T^2) && (5)\\ &= \sigma_T^2 && (6) \end{aligned}$$

by independence and the fact that U,V and T are all mean 0.

For the variance we have $Var(U+T) = 1+\sigma_T^2$ (variances add when the variables are independent) and $Var(V+T) = 1+\sigma_T^2$.

Therefore

$$Cor(X,Y) = \frac{\sigma_T^2}{1+\sigma_T^2}.$$

You can now see why the inclusion of the shared component T, makes X and Y positively correlated, and the formula for the correlation of X and Y shows that as the shared term T gets more dominant (that is as σ_T^2 increases) so the correlation goes to 1.

1.4 R simulation

We will use $\sigma_T^2 = 4$. The formula for the correlation between X and Y implies:

$$Cor(X,Y) = \frac{\sigma_T^2}{1+\sigma_T^2} = \frac{4}{1+4} = 0.8.$$

Create 10,000 reps for the simulation:

```
set.seed(12345678)
U <- rnorm(10000,0,1)
V <- rnorm(10000,0,1)
T <- rnorm(10000,0,2) #rnorm requires the sd, 2, not the variance, 4.
X <- U + T
Y <- V + T
print(cor(X,Y))
```

From the simulation $Cor(X,Y) = 0.805$ very similar to the theoretical 0.8.