STATISTICAL RETHINKING WINTER 2020/2021
HOMEWORK, WEEK 9 SOLUTIONS

1. Loading the data and running the model:

```
library(rethinking)
data(bangladesh)
d <- bangladesh

dat_list <- list(
    C = d$use.contraception,
    did = as.integer( as.factor(d$district) ),
    urban = d$urban
)

m1.1 <- ulam(
    alist(
        C ~ bernoulli( p ),
        logit(p) <- a[did] + b[did]*urban,
        c(a,b)[did] ~ multi_normal( c(abar,bbar) , Rho , Sigma ),
        abar ~ normal(0,1),
        bbar ~ normal(0,0.5),
        Rho ~ lkj_corr(2),
        Sigma ~ exponential(1)
    ) , data=dat_list , chains=4 , cores=4 , cmdstan=TRUE )
```

This is a conventional varying slopes model, with a centered parameterization. No surprises. If you peek at the posterior distributions for the average effects, you'll see that the average slope is positive:

```
precis(m1.1)
```

```
      mean   sd  5.5% 94.5% n_eff Rhat4
abar -0.69 0.10 -0.84 -0.54  1717     1
bbar  0.65 0.16  0.39  0.91   957     1
```

This implies that urban areas use contraception more. Not surprising. Now consider the distribution of varying effects:

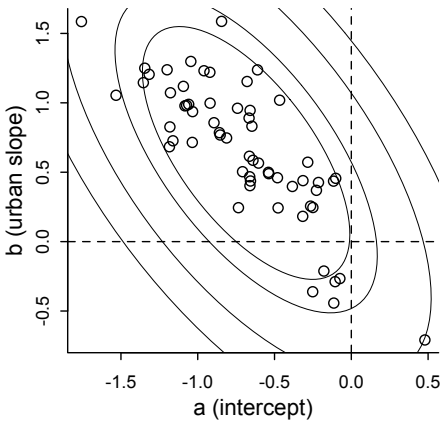
```
precis( m1.1 , depth=3 , pars=c("Rho","Sigma") )
```

```
          mean   sd  5.5% 94.5% n_eff Rhat4
Rho[1,1]  1.00 0.00  1.00  1.00   NaN   NaN
Rho[1,2] -0.65 0.17 -0.87 -0.35   437  1.00
```

```
Rho[2,1] -0.65 0.17 -0.87 -0.35   437  1.00
Rho[2,2]  1.00 0.00  1.00  1.00   NaN   NaN
Sigma[1]  0.58 0.10  0.44  0.74   737  1.00
Sigma[2]  0.79 0.20  0.49  1.12   231  1.02
```

The correlation between the intercepts and slopes is quite negative. Let's plot the individual effects to appreciate this:

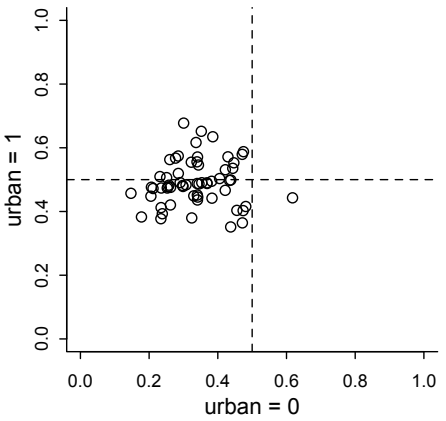
```
post <- extract.samples(m1.1)

a <- apply( post$a , 2 , mean )
b <- apply( post$b , 2 , mean )

plot( a, b , xlab="a (intercept)" , ylab="b (urban slope)" )
abline( h=0 , lty=2 )
abline( v=0 , lty=2 )

library(ellipse)
R <- apply( post$Rho , 2:3 , mean )
s <- apply( post$Sigma , 2 , mean )
S <- diag(s) %*% R %*% diag(s)
ll <- c( 0.5 , 0.67 , 0.89 , 0.97 )
for ( l in ll ) {
    el <- ellipse( S , centre=c( mean(post$abar) , mean(post$bbar) ) , level=l )
    lines( el , col="black" , lwd=0.5 )
}
```



There's the negative correlation—districts with higher use outside urban areas (a values) have smaller slopes. Since the slope is the difference between urban and non-urban areas, you can see this as saying that districts with high use in rural areas have urban areas that aren't as different.

On the outcome scale, what this ends up meaning is that urban places are much the same in all districts, but rural areas vary a lot. Plotting now in the outcome scale:

```
u0 <- inv_logit( a )
u1 <- inv_logit( a + b )

plot( u0 , u1 , xlim=c(0,1) , ylim=c(0,1) , xlab="urban = 0" , ylab="urban = 1" )
abline( h=0.5 , lty=2 )
abline( v=0.5 , lty=2 )
```



This plot is on the probability scale. The horizontal axis is probability of contraceptive use in rural area of a district. The vertical is the probability in urban area of same district. The urban areas all straddle 0.5. Most the of the rural areas are below 0.5. The negative correlation between the intercepts and slopes is necessary to encode this pattern.

In fact, if we fit the model so it instead has two intercepts, one for rural and one for urban, there is no strong correlation between those intercepts. Here's such a model:

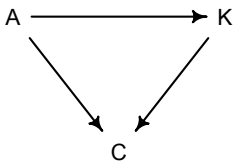
```
# version with matrix instead of slopes
dat_list$uid <- dat_list$urban + 1L

m1.2 <- ulam(
    alist(
        C ~ bernoulli( p ),
        logit(p) <- a[did,uid],
        vector[2]:a[did] ~ multi_normal( c(abar,bbar) , Rho , Sigma ),
        abar ~ normal(0,1),
        bbar ~ normal(0,1),
        Rho ~ lkj_corr(2),
```

```
        Sigma ~ exponential(1)
    ) , data=dat_list , chains=4 , cores=4, cmdstan=TRUE )
precis( m1.2 , depth=3 , pars="Rho" )
```

```
          mean   sd  5.5% 94.5% n_eff Rhat4
Rho[1,1]  1.00 0.00  1.00  1.00   NaN   NaN
Rho[1,2] -0.09 0.27 -0.51  0.36   318  1.01
Rho[2,1] -0.09 0.27 -0.51  0.36   318  1.01
Rho[2,2]  1.00 0.00  1.00  1.00   NaN   NaN
```

Correlation all gone.

2. Here's my DAG:



A is age, K is number of children, and C is contraception use. To study this DAG, we should estimate both the total causal influence of A and then condition also on K and see if the direct influence of A is smaller. Here's the model for the total influence of A:

```
dat_list$children <- standardize( d$living.children )
dat_list$age <- standardize( d$age.centered )

m2.1 <- ulam(
    alist(
        C ~ bernoulli( p ),
        logit(p) <- a[did] + b[did]*urban + bA*age,
        c(a,b)[did] ~ multi_normal( c(abar,bbar) , Rho , Sigma ),
        abar ~ normal(0,1),
        c(bbar,bA) ~ normal(0,0.5),
        Rho ~ lkj_corr(2),
        Sigma ~ exponential(1)
    ) , data=dat_list , chains=4 , cores=4 , cmdstan=TRUE )

precis(m2.1)
```

```
      mean   sd  5.5% 94.5% n_eff Rhat4
abar -0.68 0.10 -0.85 -0.53  1714     1
bA    0.08 0.05  0.00  0.16  3927     1
bbar  0.64 0.16  0.38  0.89  1247     1
```

In this model, the total causal effect of age is positive and very small. Older individuals use slightly more contraception.

And now the model with both K and A:

```
m2.2 <- ulam(
    alist(
        C ~ bernoulli( p ),
        logit(p) <- a[did] + b[did]*urban + bK*children + bA*age,
        c(a,b)[did] ~ multi_normal( c(abar,bbar) , Rho , Sigma ),
        abar ~ normal(0,1),
        c(bbar,bK,bA) ~ normal(0,0.5),
        Rho ~ lkj_corr(2),
        Sigma ~ exponential(1)
    ) , data=dat_list , chains=4 , cores=4 , cmdstan=TRUE )

precis(m2.2)
```

```
      mean   sd  5.5% 94.5% n_eff Rhat4
abar -0.72 0.10 -0.88 -0.55  1387     1
bA   -0.26 0.07 -0.38 -0.15  2222     1
bK    0.51 0.07  0.40  0.62  1977     1
bbar  0.69 0.16  0.43  0.94   850     1
```

In this model, the direct effect of age is negative, and much farther from zero than before. The effect of number of children is strong and positive. These results are consistent with the DAG, because they imply that the reason the total effect of age, from `m2.1`, is positive is that older individuals also have more kids. Having more kids increases contraception. Being older, controlling for kids, actually makes contraception less likely.

3. To build this model, you need the ordered categorical predictor machinery from the book example. The maximum observed number of kids in the sample is 4. So that means we need three parameters, for three transitions in number of kids. We'll set up the alpha prior that way:

```
dat_list$K <- d$living.children
dat_list$alpha <- rep(2,3)

m3.1 <- ulam(
    alist(
        C ~ bernoulli( p ),
        logit(p) <- a[did] + b[did]*urban + bK*sum( delta_shell[1:K] ) + bA*age,
        c(a,b)[did] ~ multi_normal( c(abar,bbar) , Rho , Sigma ),
        abar ~ normal(0,1),
        c(bbar,bK,bA) ~ normal(0,0.5),
        Rho ~ lkj_corr(2),
        Sigma ~ exponential(1),
        vector[4]: delta_shell <<- append_row( 0 , delta ),
        simplex[3]: delta ~ dirichlet( alpha )
    ) , data=dat_list , chains=4 , cores=4 , cmdstan=TRUE )
```

```
precis(m3.1)
```

```
      mean   sd  5.5% 94.5% n_eff Rhat4
abar -1.56 0.15 -1.80 -1.33   503  1.01
bA   -0.23 0.06 -0.33 -0.13  1028  1.00
bK    1.27 0.15  1.03  1.51   554  1.01
bbar  0.69 0.16  0.43  0.95   947  1.00
```

So the general effects are the same—age reduces use and kids increase it. Let's look at the individual kid parameters now:

```
precis( m3.1 , 3 , pars="delta" )
```

```
         mean   sd 5.5% 94.5% n_eff Rhat4
delta[1] 0.74 0.08 0.60  0.87  2235     1
delta[2] 0.16 0.08 0.05  0.30  2411     1
delta[3] 0.09 0.06 0.02  0.19  3247     1
```

`delta[1]` is the transition from 1 to 2 kids. It is much larger than the other two parameters. So most of the influence of kids on contraception comes from having a second child.