Homework 8

Question 1

Revisit the Reed frog survival data, data(reedfrogs), and add the predation and size treatment variables to the varying intercepts model. Consider models with either predictor alone, both predictors, as well as a model including their interaction. What do you infer about the causal influence of these predictor variables? Also focus on the inferred variation across tanks (the across tanks). Explain why it changes as it does across models with different predictors included.

```
data(reedfrogs)

d_q1 <- reedfrogs

head(d_q1)
```

```
##   density pred  size surv propsurv
## 1      10   no   big    9      0.9
## 2      10   no   big   10      1.0
## 3      10   no   big    7      0.7
## 4      10   no   big   10      1.0
## 5      10   no small    9      0.9
## 6      10   no small    9      0.9
```

```
dat_list_q1 <- list(
  S=d_q1$surv,
  D=d_q1$density,
  P=as.integer(d_q1$pred) - 1L,
  Z=as.integer(d_q1$size),
  N=48
)
```

First model surv with only pred alone

```
model_q1_1_fit <- stan(file='week08/08_q1_1.stan', data=dat_list_q1, cores=4)
```

```
## Warning in readLines(file, warn = TRUE): incomplete final
## line found on 'C:\Users\Orcun Gumus\OneDrive - McKinsey &
## Company\Desktop\statrethinking_winter2019\week08\08_q1_1.stan'
```

```
model_q1_2_fit <- stan(file='week08/08_q1_2.stan', data=dat_list_q1, cores=4)
```

```
model_q1_3_fit <- stan(file='week08/08_q1_3.stan', data=dat_list_q1, cores=4)
```

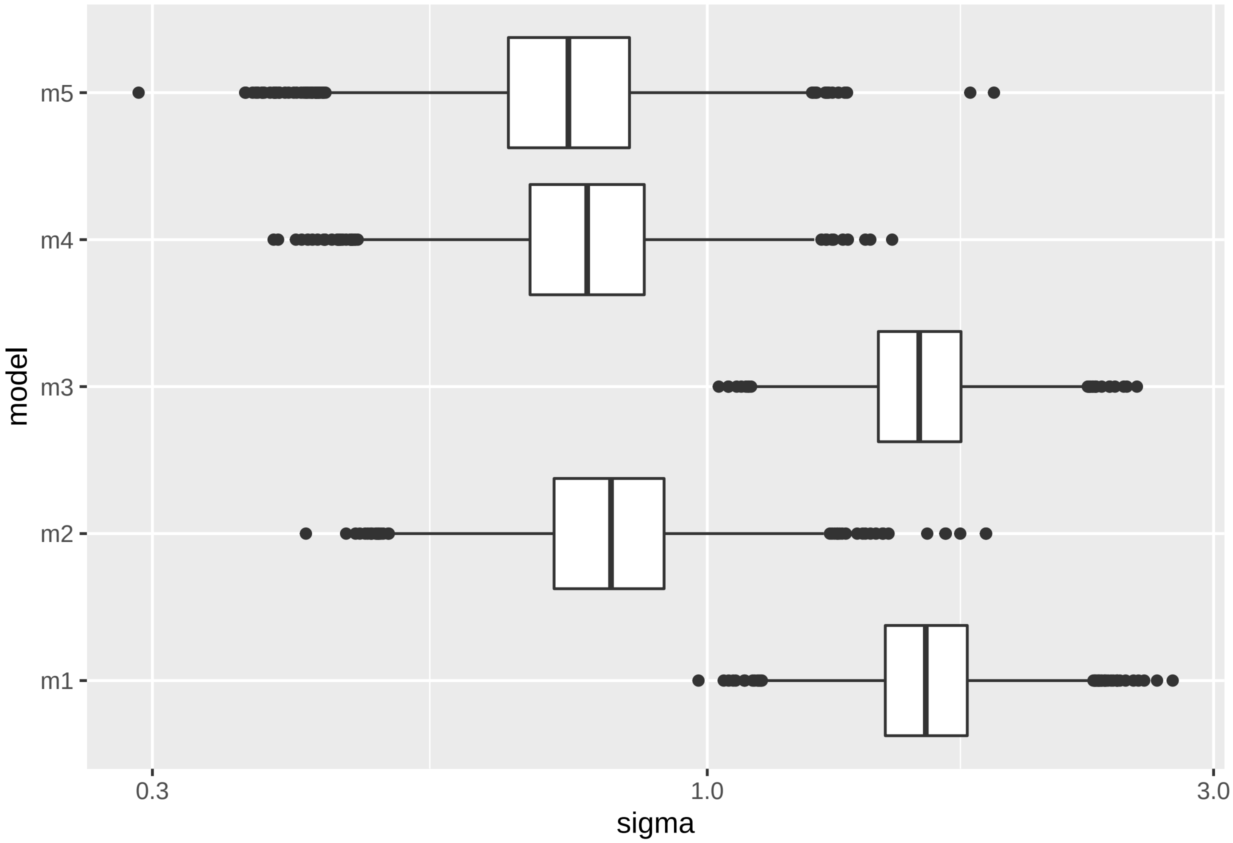
```
model_q1_4_fit <- stan(file='week08/08_q1_4.stan', data=dat_list_q1, cores=4)
```

```
model_q1_5_fit <- stan(file='week08/08_q1_5.stan', data=dat_list_q1, cores=4)
```

```
## Warning in readLines(file, warn = TRUE): incomplete final
## line found on 'C:\Users\Orcun Gumus\OneDrive - McKinsey &
## Company\Desktop\statrethinking_winter2019\week08\08_q1_5.stan'
```

```
results <- rbind(
  data.frame(model='m5', sigma=rstan::extract(model_q1_5_fit)$A_sigma),
  data.frame(model='m4', sigma=rstan::extract(model_q1_4_fit)$A_sigma),
  data.frame(model='m3', sigma=rstan::extract(model_q1_3_fit)$A_sigma),
  data.frame(model='m2', sigma=rstan::extract(model_q1_2_fit)$A_sigma),
  data.frame(model='m1', sigma=rstan::extract(model_q1_1_fit)$A_sigma)
)
```

```
results %>%
  ggplot(aes(x=model, y=sigma)) +
  geom_boxplot()+
  coord_flip()+
  scale_y_log10()
```



Question 2

Now, focus on predicting use.contraception, clustered by district_id. Fit both (1) a traditional fixed-effects model that uses an index variable for district and (2) a multilevel model with varying intercepts for district. Plot the predicted proportions of women in each district using contraception, for both the fixed-effects model and the varying-effects model. That is, make a plot in which district ID is on the horizontal axis and expected proportion using contraception is on the vertical. Make one plot for each model, or layer them on the same plot, as you prefer. How do the models disagree? Can you explain the pattern of disagreement? In particular, can you explain the most extreme cases of disagreement, both why they happen where they do and why the models reach different inferences?

```
data(bangladesh)

d_q2 <- bangladesh

d_q2$district_id <- as.integer(as.factor(d_q2$district))

dat_list_q2 <- list(
  C=d_q2$use.contraception,
  D=d_q2$district_id,
  N=1934,
  K=max(d_q2$district_id)
)
```

```
model_q2_1_fit <- stan(file='week08/08_q2_1.stan', data=dat_list_q2, cores=4)
```

```
## Warning in readLines(file, warn = TRUE): incomplete final
## line found on 'C:\Users\Orcun Gumus\OneDrive - McKinsey &
## Company\Desktop\statrethinking_winter2019\week08\08_q2_1.stan'
```

```
model_q2_2_fit <- stan(file='week08/08_q2_2.stan', data=dat_list_q2, cores=4)
```

```
## Warning in readLines(file, warn = TRUE): incomplete final
## line found on 'C:\Users\Orcun Gumus\OneDrive - McKinsey &
## Company\Desktop\statrethinking_winter2019\week08\08_q2_2.stan'
```

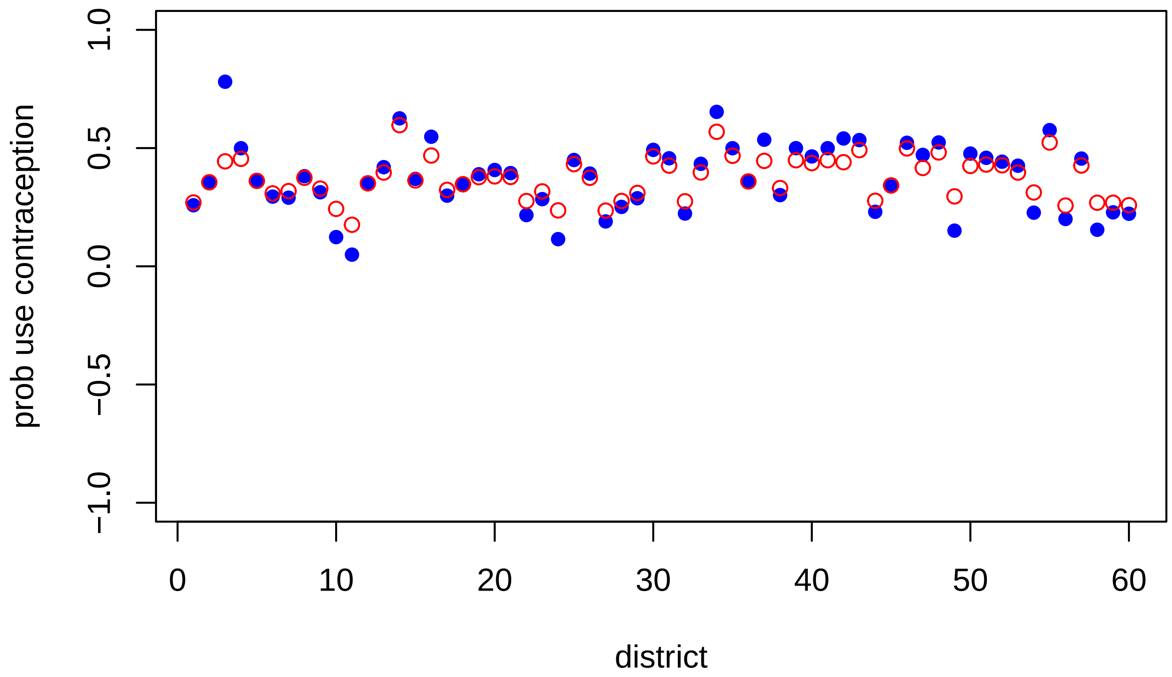
```
compare(model_q2_1_fit, model_q2_2_fit)
```

```
##                     WAIC       SE    dWAIC     dSE    pWAIC     weight
## model_q2_2_fit 2514.674 24.98381 0.000000      NA 35.72279 0.98983926
## model_q2_1_fit 2523.832 28.89257 9.158022 7.70615 53.88318 0.01016074
```

```
post1 <- extract.samples( model_q2_1_fit )
post2 <- extract.samples( model_q2_2_fit )

p1 <- apply( post1$dD, 2 , mean )
p2 <- apply( post2$dDk, 2 , mean )

nd <- max(dat_list_q2$D)
plot( NULL , xlim=c(1,nd) , ylim=c(-1,1) , ylab="prob use contraception" ,
xlab="district" )
points( 1:nd , inv_logit(p1) , pch=16 , col='blue' )
points( 1:nd , inv_logit(p2), col='red')
```



```
#abline( h=mean(inv_logit(post2$a_bar)) , lty=2 )
```

Question 3

Return to the Trolley data, data(Trolley), from Chapter 12. Define and fit a varying intercepts model for these data. By this I mean to add an intercept parameter for the individual to the linear model. Cluster the varying intercepts on individual participants, as indicated by the unique values in the id variable. Include action, intention, and contact as before. Compare the varying intercepts model and a model that ignores individuals, using both WAIC/LOO and posterior predictions. What is the impact of individual variation in these data?

```
data(Trolley)

d_q3 <- Trolley

edu_levels <- c(6, 1, 8, 4, 7, 2, 5, 3)
d_q3$edu_new <- edu_levels[d_q3$edu]

dat_list_q3 <- list(
  response=d_q3$response,
  P=as.numeric(factor(d_q3$id)),
  A=d_q3$action,
  I=d_q3$intention,
  C=d_q3$contact,
  IA=d_q3$intention * d_q3$action,
  IC=d_q3$intention * d_q3$contact,
  E=as.integer(d_q3$edu_new),
  F=1-d_q3$male,
  AGE=standardize(d_q3$age),
  K=7,
  NP=331,
  N=9930
)
```

```
model_q3_1_fit <- stan(file='week08/08_q3_1.stan', data=dat_list_q3, cores=4)
```

```
## Warning in readLines(file, warn = TRUE): incomplete final
## line found on 'C:\Users\Orcun Gumus\OneDrive - McKinsey &
## Company\Desktop\statrethinking_winter2019\week08\08_q3_1.stan'
```

```
model_q3_2_fit <- stan(file='week08/08_q3_2.stan', data=dat_list_q3, cores=4)
```

```
## Warning in readLines(file, warn = TRUE): incomplete final
## line found on 'C:\Users\Orcun Gumus\OneDrive - McKinsey &
## Company\Desktop\statrethinking_winter2019\week08\08_q3_2.stan'

## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess

## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quant
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#tail-ess
```

```
loo(model_q3_1_fit)
```

```
##
## Computed from 4000 by 9930 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo -18331.7 43.6
## p_loo        16.7  0.1
## looic     36663.3 87.2
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
loo(model_q3_2_fit)
```

```
##
## Computed from 4000 by 9930 log-likelihood matrix
##
##          Estimate    SE
## elpd_loo -15530.7  89.9
## p_loo       359.5   4.7
## looic     31061.4 179.7
## ------
## Monte Carlo SE of elpd_loo is 0.2.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
compare(model_q3_1_fit, model_q3_2_fit)
```

```
##                     WAIC        SE    dWAIC      dSE     pWAIC weight
## model_q3_2_fit 31059.76 179.72455    0.000       NA 358.72347      1
## model_q3_1_fit 36663.30  87.22013 5603.534 173.7226  16.68227      0
```