

STATISTICAL RETHINKING 2025

WEEK 6 SOLUTIONS

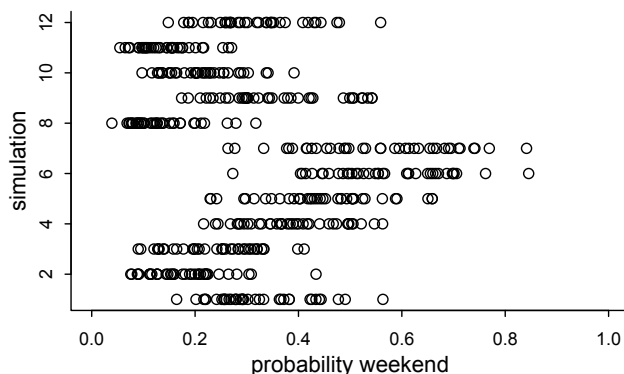
1. The model we need is very similar to the no-pooling models already seen in the course. As an `ulam` model:

```
dat <- read.csv("BMJSubmissions.csv")
m1 <- ulam(
  alist(
    W ~ bernoulli(p),
    logit(p) <- a + b[L],
    a ~ normal(-1,1),
    b[L] ~ normal(0,0.5)
  ), data=dat , chains=4 , cores=4 )
```

Before going further, let's think about the priors. What is a sensible prior for the probability of submitting on the weekend? We can start by simulating from these priors, to see what they imply.

```
n <- 12
plot(NULL,xlim=c(0,1),ylim=c(1,n),xlab="probability weekend",
     ylab="simulation")

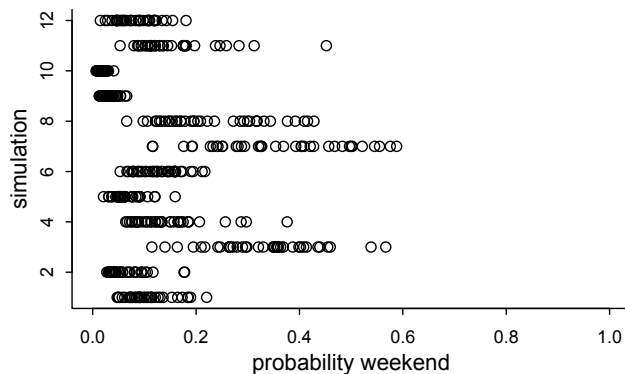
for ( i in 1:n ) {
  # average log-odds of submitting on weekend
  a <- rnorm(1,-1,1)
  # offset for each country
  b <- rnorm(38,0,0.5)
  # probabilities for each country
  p <- inv_logit(a + b)
  points( p , rep(i,38) )
}
```



The rows are independent simulations from the prior, and each point is a simulated country probability. So right away you can see that these priors allow countries to vary a lot. And the average also varies a lot. Are these priors good? Well, what do you think is a good guess for the probability of submitting on the weekend? If submission were random, then the probability should be $1/7$, which is about 0.14. In that light, these priors are way too high on average. They routinely have countries preferring to submit on weekends. But we do want to allow for probabilities above 0.14, because maybe there is sometimes a tendency to do such work on the weekend. Maybe a maximum of 0.5 would make more sense? In that case we might try instead:

```
m1 <- ulam(
  alist(
    W ~ bernoulli(p),
    logit(p) <- a + b[L],
    a ~ normal(-2,1),
    b[L] ~ normal(0,0.5)
  ), data=dat , chains=4 , cores=4 )
```

Modify the prior simulation code to reflect these new priors, and you will get a plot like:



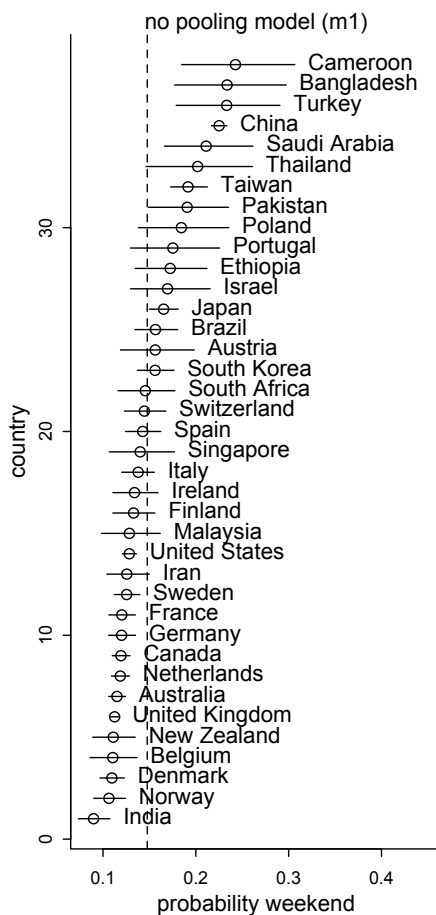
The average is a lot lower now, but high averages are still plausible, and countries can vary a lot even when the mean is low. I encourage you to play around and get a feel for how the mean and variation interact in the outcome space (probability instead of log-odds).

With model `m1` fit, check the chains. They should look good.

Now let's visualize the estimates. I will use `link` to compute posterior predictions for each country.

```
pred1 <- link(m1,data=list(L=1:38))
post1 <- extract.samples(m1)

mus <- apply(pred1,2,mean)
o <- order(mus)
plot( mus[o] , 1:38 , xlab="probability weekend" , ylab="country" ,
      xlim=c(0.08,0.45) )
cis <- apply(pred1,2,PI)
for ( i in 1:38 ) lines( cis[,o[i]] , c(i,i) )
for ( i in 1:38 ) text( labels=levels(country)[o[i]] , x=cis[2,o[i]] ,
                      y=i , pos=4 )
abline( v=mean(inv_logit(post1$a)) , lty=2 )
mtext("no pooling model (m1)")
```



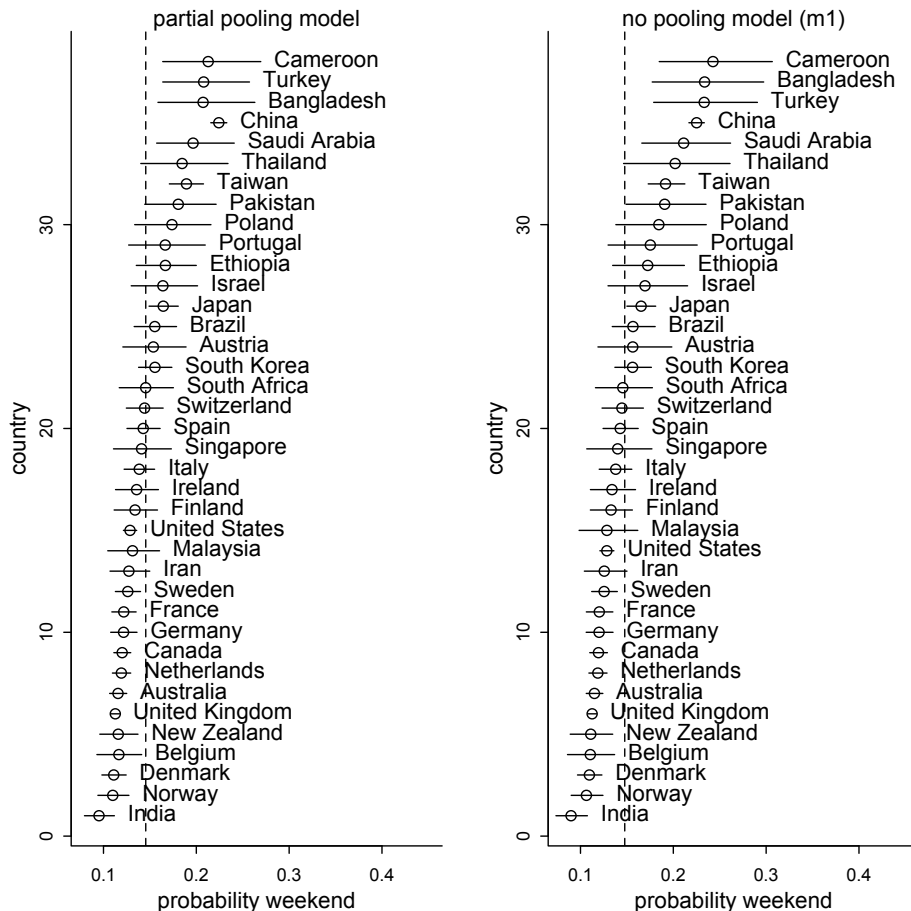
The vertical dashed line is the average across nations. Cameroon, Bangladesh, and Turkey are on top, but they have wide uncertainty intervals. China is precisely estimated to be almost the same. At the bottom, India is the clear winner, followed by some Nordic countries, and a variety of other European and quasi-EU nations.

2. For the partial-pooling model, we replace the fixed standard deviation in the prior for $b[L]$ with a parameter σ . Like this:

```
m2 <- ulam(
  alist(
    W ~ bernoulli(p),
    logit(p) <- a + b[L],
    a ~ normal(-2,1),
    b[L] ~ normal(0,sigma),
    sigma ~ exponential(1)
  ), data=dat , chains=4 , cores=4 )
```

This model should mix fine. Check the chain diagnostics and trace/trunk plots so you get used to interpreting them.

Below I plot the new estimates on the left, with the previous no-pooling on right for comparison.



Notice that the top three have shifted left a bit, towards the mean. This is because there is less data for those nations, so they have wider intervals, so they are influenced more by the global average in the partial-pooling model. Almost every nation with a wide interval has shifted towards the mean, in the left plot.

We can visualize this by plotting the log sample size in each country against the change in the posterior mean after applying partial-pooling. Like this:

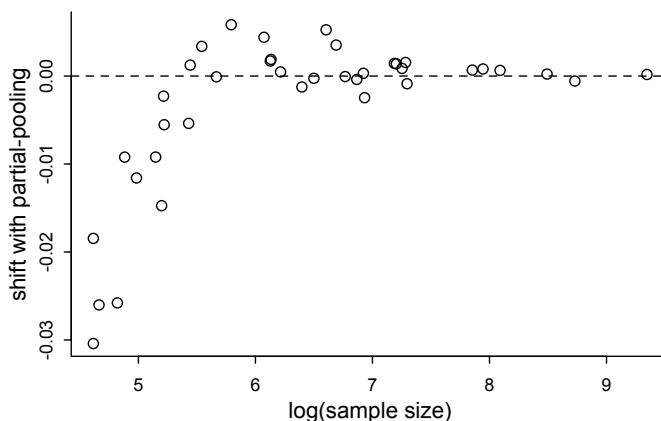
```
mu1 <- apply(pred1,2,mean)
mu2 <- apply(pred2,2,mean)
shift <- mu2 - mu1

# shift against sample size
```

```

N <- table(dat$L) # sample size for each nation
plot( log(as.numeric(N)) , shift , xlab="log(sample size)" ,
      ylab="shift with partial-pooling" )
abline(h=0,lty=2)

```



The nation on the left have smaller samples sizes (fewer submissions). Most of the em have gotten smaller after partial-pooling. Those on the right have very large sample sizes. Partial-pooling had much less of an effect, if any. This is what you should expect, if the model does what you want it to do.

Note however that those nations with smaller sample sizes could really be more different from the mean. There just isn't enough evidence in the sample. So they get shrunk towards the mean. It is interesting after all that most of the nations with small sample sizes are above the mean, not below it.