Forecasting Demo

Josh Tyler

Introduction

This is an example analysis report for the mpox forecasting competition. Here we outline an analysis workflow that starts at loading in the data and ends with the creation of the Estimates in the correct submission format.

Organising The Data

In order to create a forecast, we must first load and clean the raw cases data. The historic data was filtered to include only months where all six WHO regions had case data provided and then a summary table was produced with each region having a total case number per month. We have also indexed the months for easier modelling and taken the log of the case numbers.

```r
library(tidyverse)
library(lubridate)
library(brms)
library(gt)
set.seed(1)

df<-read.csv(file = "data/aggregated_mpox_case_data.csv") |>
  as.data.frame() |>
  filter(month_start>"2022-06-01")|>
  group_by(who_region,month_start) |>
  summarise(cases=sum(cases)) |>
  as.data.frame()

df$cases<-log10(df$cases)

table(df$month_start)
```

```
df$month_start<-as.Date(df$month_start)

date_lookup<-data.frame(month_start=unique(df$month_start),
                        index=1:length(unique(df$month_start)))

who_regions<-unique(df$who_region)


df<-merge(x = df,y = date_lookup,by = "month_start") |>
  arrange(index,who_region)
df$m<-paste0(df$who_region,df$index)

end_index<-tail(unique(df$index),1)
```

Plotting The Data

It is always important to check that the original data looks correct so we have replicated the figures from the WHO dashboard sowing case numbers per month for each of the 6 regions.

```
ggplot(data = df)+
  geom_bar(aes(x=month_start,y=cases,fill = who_region),stat = "identity",col="black")+
  theme_bw()+
  facet_grid(who_region~.,scales = "free_y")+ scale_fill_discrete(name = "WHO Region")+
  xlab("Date")
```
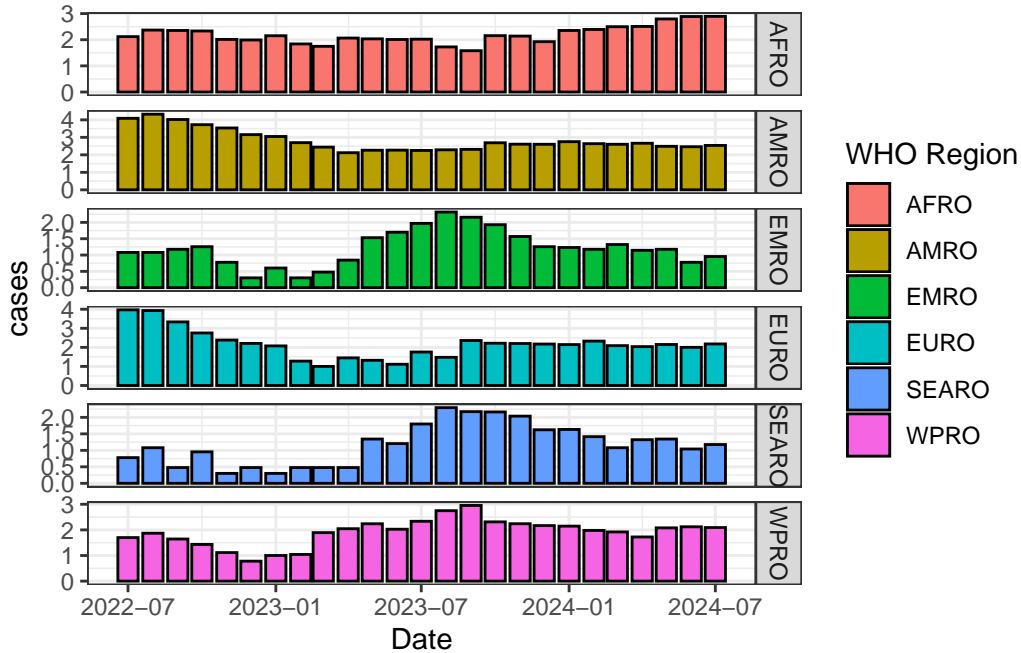
Creating a Forecast

In this example, we have chosen to use the `brms` package to fit a Bayesian Regression model, which is then extended to the remainder of 2024.

```
fit1<-brm(cases ~ s(index,by=who_region,bs="bs"),data = df,
          family = gaussian(),cores = 8,chains = 8,iter = 5000)

newdf <- expand.grid(who_region=who_regions,index=1:(end_index+5))
newdf <- cbind(newdf, fitted(fit1, newdata = newdf)) |>
  arrange(index,who_region)
names(newdf) <- make.names(names(newdf))
newdf$m<-paste0(newdf$who_region,newdf$index)

tmp<-df[,c(3,5)]

newdf<-merge(x = newdf,y=tmp,by="m",all = T) |>
  arrange(index,who_region)
```

We can see that the fits are broadly in keeping with the historic data and that the prediction envelope expands rapidly as we move to the remaining months of 2024. This is due to our

analyses being a "simple" regression rather than taking advantage of other datasets to help guide the predictions.

```
ggplot(data = newdf,aes(x=index)) +
    geom_bar(aes(y=cases,fill = who_region),stat = "identity",col="black")+
    geom_smooth(aes(y=Estimate,ymin = Q2.5, ymax = Q97.5),stat = "identity")+
  facet_grid(who_region~.,scales = "free_y")+
  theme_bw()
```
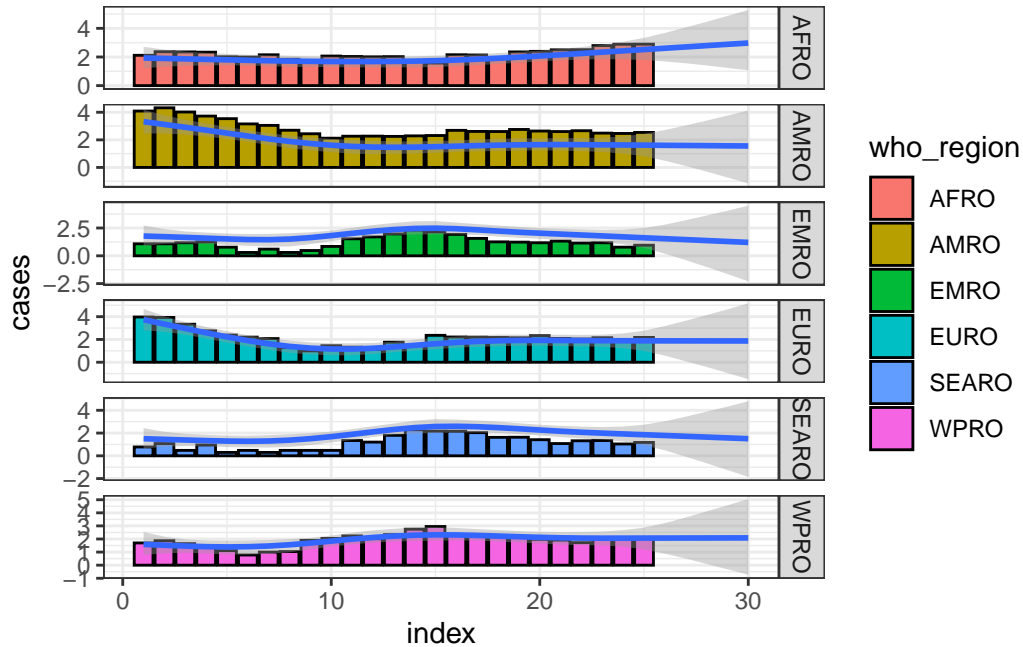

Creating the results table

The estimates of this competition must be given in a csv with 3 columns: WHO region, month_start and Estimate. Here we filter our estimates and convert the log(cases) to actual estimates and then save the results to be considered.

```
results<-newdf |>
  filter(index>25) |>
  select(who_region,index,Estimate) |>
  mutate(Estimate=floor(10^Estimate))

results<- results |>
```

```
  mutate(month_start=as.Date(df$month_start[1]) %m+% months(results$index-1)) |>
  select(who_region,month_start,Estimate)

write.csv(results,"Mpox_forecast.csv",row.names = F)
```

```
gt(results)
```

who_region	month_start	Estimate
AFRO	19936	408
AMRO	19936	40
EMRO	19936	34
EURO	19936	75
SEARO	19936	60
WPRO	19936	117
AFRO	19967	503
AMRO	19967	39
EMRO	19967	28
EURO	19967	74
SEARO	19967	51
WPRO	19967	118
AFRO	19997	620
AMRO	19997	37
EMRO	19997	23
EURO	19997	74
SEARO	19997	44
WPRO	19997	119
AFRO	20028	765
AMRO	20028	36
EMRO	20028	19
EURO	20028	74
SEARO	20028	37
WPRO	20028	120
AFRO	20058	943
AMRO	20058	35
EMRO	20058	16
EURO	20058	73
SEARO	20058	31
WPRO	20058	121