ADS2 Group Exercise

Group 3

2021/5/6

We used several figures in the document. We want to label the figures as "Figure a.b", where "a" means which part the figure belongs to and "b" means the order of the figure. "Figure 1.1" means the first figure in part 1.

Import necessary libraries

```
library("ggplot2")
library("ggpubr")
library("knitr")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Part 1: Exploring the data

import data

```
data <- read.csv("C:/Users/16977/Desktop/ADS ICA/substance_use.csv")
```

In 2019, what region of the world has the highest rate of alcohol-related deaths among men aged 40-44?

```
Q1_1 <- data[which(data$measure=="Deaths"
                   & data$year==2019
                   & data$cause=="Alcohol use disorders"
                   & data$age=="40 to 44"
                   & data$sex=="Male"),]
a <- max(Q1_1$val)
res <- Q1_1[which(Q1_1$val==a),"location"]
res
```

```
## [1] "Europe & Central Asia - WB"
```

Europe & Central Asia has the highest rate of alcohol-related deaths among men aged 40-44.

Looking at the prevalence of alcohol-related disease in the East Asia and Pacific region, how has this changed over time and in the different age groups (data.frame: Q1_2_1)? Is there a difference between men and women (data.frame: Q1_2)?

Extract the data that is necessary for this question.

```
Q1_2 <- data[which(data$measure=="Prevalence"
                   & data$cause=="Alcohol use disorders"
                   & data$location=="East Asia & Pacific - WB"),]
```

Q1_2: this is a data frame which contains the prevalence of alcohol use disorders in east Asia & pacific. (female and male separated)

The first part of the question (how has this changed over time and in the different age groups?)

The first part of the question does not ask us about the gender. Given that within a given region and age group, the size of the gender groups is approximately identical, we want to take the average of the prevalence of male and female.

```
Q1_2_1 <- aggregate(val~age+year,Q1_2,FUN=sum)
```

Q1_2_1: this is a data frame which contains the prevalence of alcohol use disorders in east Asia & pacific. (female and male combined-take average)

We plot the data to see how it looks like.

```
h <- ggplot(Q1_2_1,aes(x=year,y=val,color=age))
h <- h + geom_line(size=1)
h <- h + geom_point()
h <- h + labs(title="The prevalence of alcohol use disorders in East Asia and Pacific
(age group combined)",
y = "Prevalence (%)",
caption = "Figure 1.1")
h
```

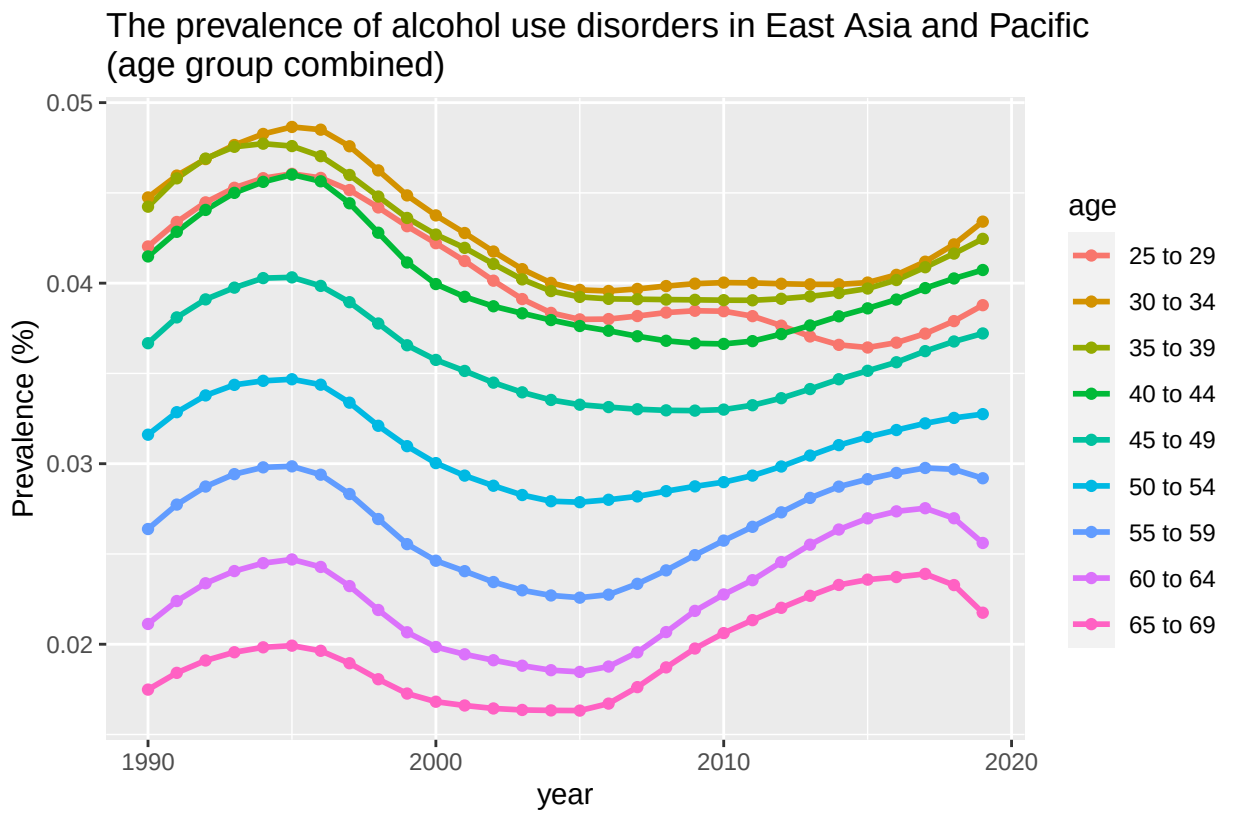

Figure 1.1

```
Q1_2_1$pattern=as.factor(c(1,1,2,2,2,2,3,3,3))
g <- ggplot(Q1_2_1,aes(x=year,y=val,col=pattern))
g <- g + facet_wrap(~age)
g <- g + geom_line(size=1)
g <- g + geom_point()
g <- g + labs(title="The prevalence of alcohol use disorders in East Asia and Pacific
(age group seperated)",
y = "Prevalence (%)",
caption = "Figure 1.2")
g <- g + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
g
```

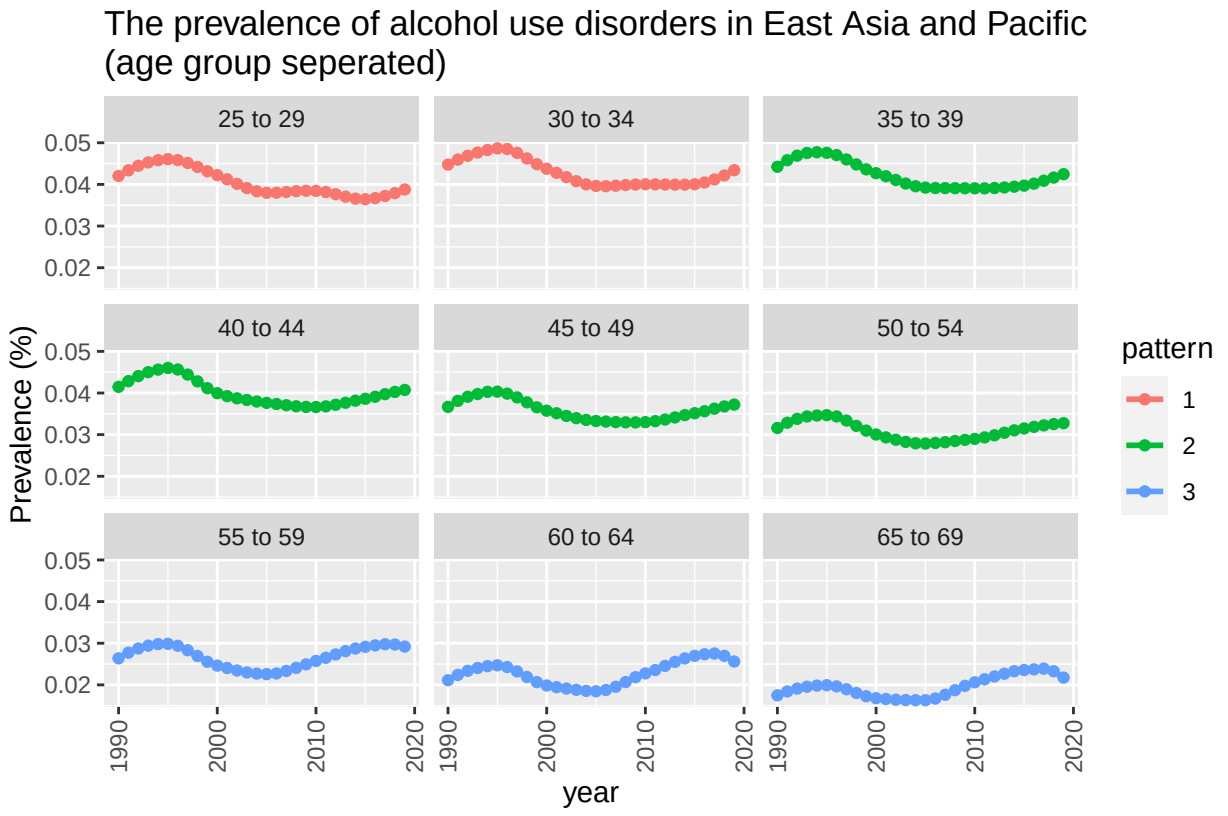

Figure 1.2

When we look at Figure 1.1 that combines the prevalence of alcohol use disorders in different age groups in East Asia and Pacific, we could identify mostly 3 patterns of changing. And it is interesting to find that the groups for the three patterns are the eldest three, four in the middle, and youngest two as shown by color in Figure 1.2.

Generally, however the prevalence changes over time, it has a peak at about 1995 and overall tendency of increase-decrease-increase. Also, except for prevalence of age group 25-29, the prevalence of a younger group is always higher than that of an elder group.

The first type of changing appears in age group 55-59, 60-64 and 65-69. The prevalence of these groups first has an increase from 1990 to approximate 1995, then appears to decrease until about 2005, followed by an increase continues until 2017, and finally there are little drops.

The second type, corresponding to age group 35-39, 40-44, 45-49, 50-54, is similar to the first type. These groups have similar trend and turning points, the difference is that in the second increase stage, the prevalence of these groups has a slower increase compared to that in the first type. Also, the little drop in the first type does not appear in the second type.

The third type relates to age group 25-29, 30-34, the two youngest groups. The first increase stage is similar, but in the decrease stage, the first decrease only continues to about 2005, then starts to increase a little bit for 5 years followed by mild decrease until about 2015, and finally appears to increase till end.

The second part of the question (Is there a difference between men and women?)

We first plot the data.

```
h <- ggplot(Q1_2,aes(x=year,y=val,color=age))
h <- h + geom_line(size=1)
h <- h + geom_point()
h <- h + labs(title="Prevalence of alcohol use disorders in East Asia & Pacific",y="Prevalence (%)",cap
h <- h + facet_grid(cols=vars(sex))
h <- h + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
h
```

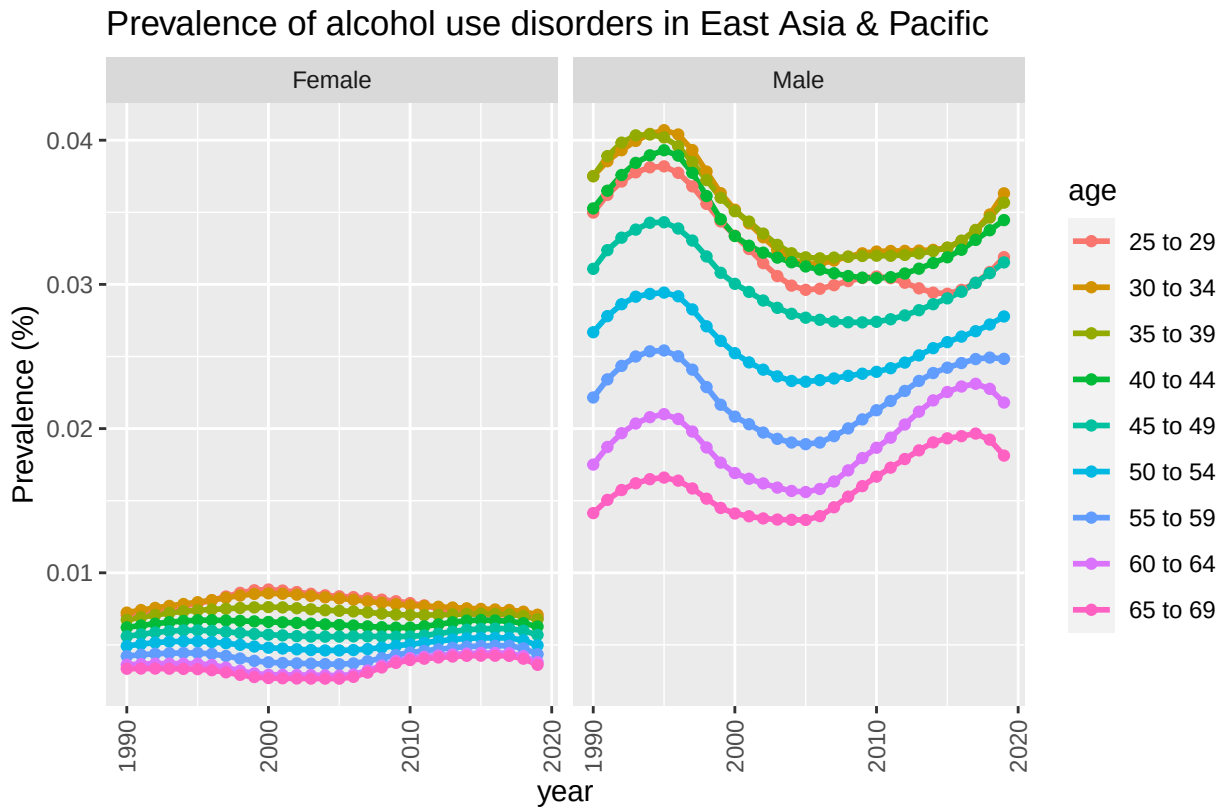

Figure 1.3

From this figure, we can see that male has larger prevalence than female in all age groups.

Generally, the prevalence decreases with age, except 25 to 29 in male. Female in all age groups have the prevalence below 0.01% with slight fluctuations during 1990 to 2020. The male in all age groups have the prevalence above 0.01%, and there is a large difference among different age groups, ranging from 0.014% to 0.04%. However, the tendency shows a similarity in different age groups of male.

```
g <- ggplot(Q1_2,aes(year,val,color=sex))
g <- g + geom_errorbar(aes(ymin=lower, ymax=upper))
g <- g + geom_line(size=1)
g <- g + geom_point(size=0.5)
g <- g + labs(title="Prevalence of alcohol use disorders in East Asia & Pacific",y="Prevalence (%)",cap
g <- g + facet_wrap(~age)
g <- g + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
g
```

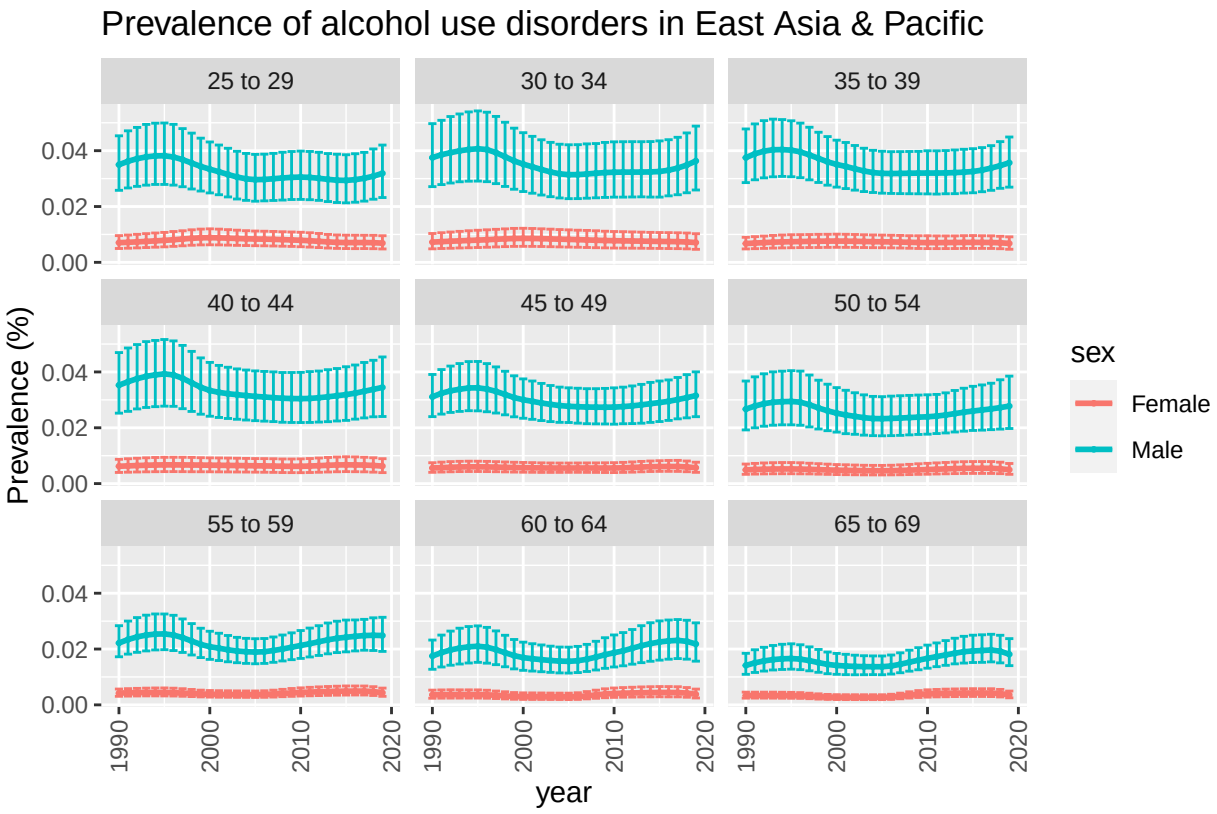

Figure 1.4

Then, we separate different age groups and plot again. We can clearly see that male has the higher prevalence than female, since there is no overlap of the confidence interval of male and female during 1990-2019.

In the United States, there is talk of an "Opioid epidemic". Part of the problem is that since the late 1990s, doctors have increasingly been prescribing pain killers which can be highly addictive. Looking at the data from the United States, can you confirm an increase in the prevalence of diseases related to opioid use? What age group is the most affected? (data.frame: Q1_3)

Our understanding to the question:

The question says doctors have increasingly been prescribing opioid **since the late 1990s**. Since we have the 1990-2019 data, we think we should separate the data into two parts (before the start of increased prescription of opioid (1990-1999) and after the start of increased prescription of opioid (2000-2019)) and compare whether there is a difference. The reason for choosing **2000** as the break point is that it takes time for the increased prescription to be reflected in the prevalence. In other words, the increased opioid prescription will not cause the disorders instantly. It may takes a few years.

First, we extract the relevant data.

```
temp <- data[which(data$location=="North America"
                   & data$cause=="Opioid use disorders"
                   & data$measure=="Prevalence"),]
```

Given that within a given region and age group, the size of the gender groups is approximately identical, we take the average of the prevalence of male and female to get the whole population data.

```
Q1_3 <- aggregate(val~age+year,temp,FUN=sum)
```

Q1_3: the data frame contains the age, year, prevalence value of Opioid use disorders in north America. (no sex information since we took the average of male and female.)

We plot the data and see how it looks like.

```
h <- ggplot(Q1_3,aes(x=year,y=val,color=age))
h <- h + geom_line(size=0.5)
h <- h + geom_point(size=1)
h <- h + labs(title = "The prevalence of opioid use disorders
in the North America (combined)",
              y = "Prevalence (%)",caption = "Figure 1.5")
h
```

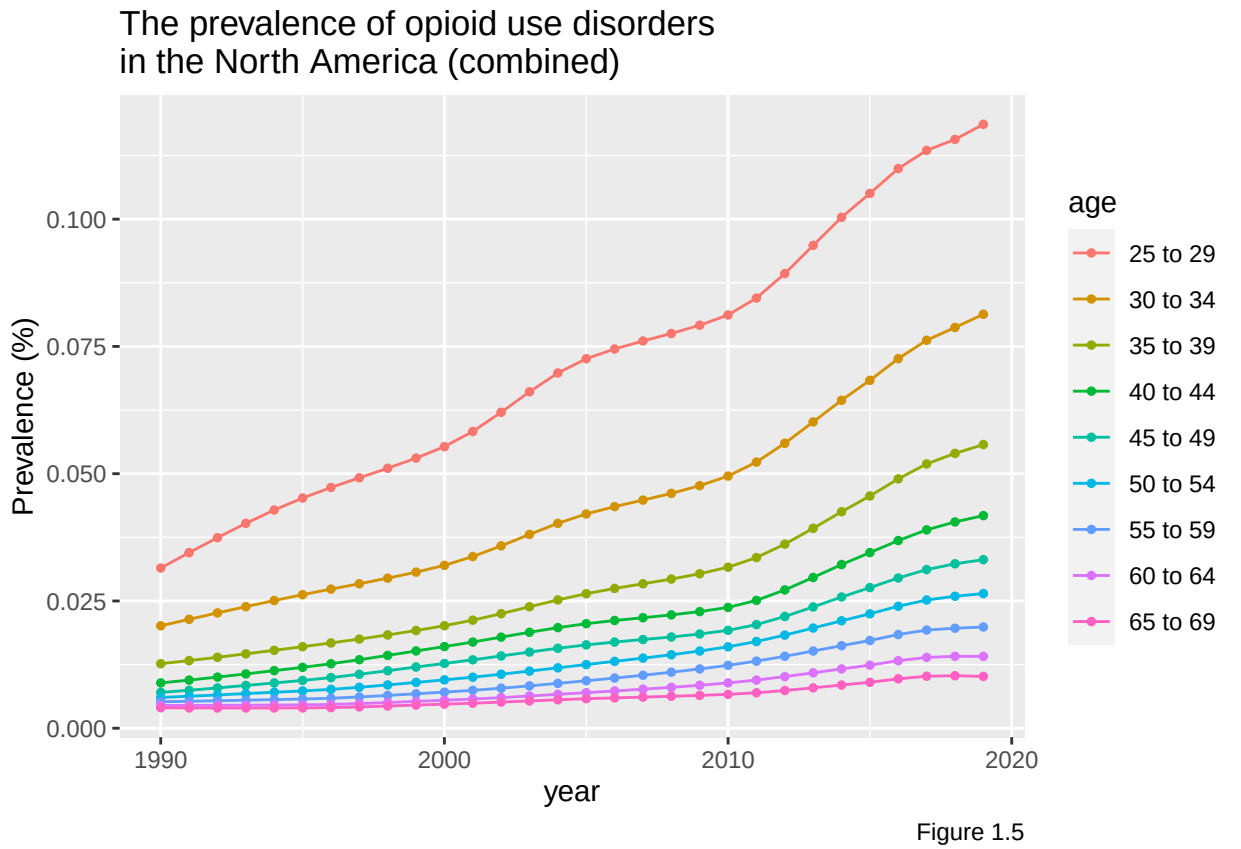

Figure 1.5

```
h <- ggplot(Q1_3,aes(x=year,y=val,color=age))
h <- h + geom_line(size=0.5)
h <- h + geom_point(size=1)
h <- h + facet_grid(cols=vars(age))
h <- h + labs(title = "The prevalence of opioid use disorders
in the North America (separated)",
              y = "Prevalence (%)",caption = "Figure 1.6")
h <- h + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
h
```

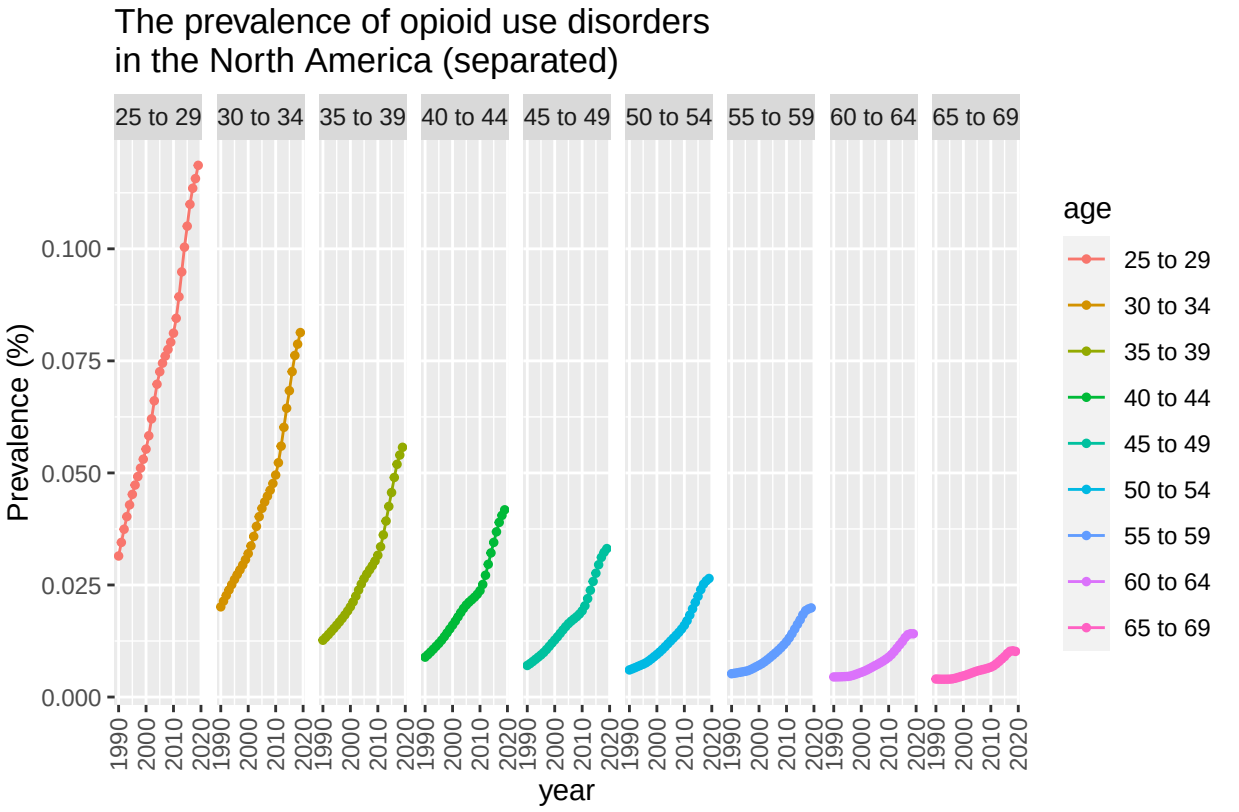

Figure 1.6

From two plots, we can see there is an increase in prevalence in all age groups.

Since the prevalence of both "before" period and "after" period is increasing, we think we should compare the speed (slope) of the increase of prevalence rather than the prevalence itself. If the prevalence of "after" period is increasing more rapidly than "before" period, we can tell that the doctors' prescription has some effect on the prevalence.

The first part of the question (can you confirm an increase..?)

To test whether there is a significance between the slope of the "before" and "after" period, we are aware that there are several ways. For example, the first possible choice is to calculate the difference of prevalence between each two years and group by "before" "after" and do tests. Another possible way is to use slope of "before" period to get predicted data points in "after" period and do tests to analyze whether they are different from real data points. After considering many possible choices, we cannot tell a significantly better way so we just chose the first one. So we will get 29 differences from 30 years data. We separated the differences according to the periods we previous mentioned. So the differences between 1990-1991, 1991-1992...1999-2000 will be grouped into "before" period while the other will be grouped into "after" period.

Since the experimental design ensured the independent random sampling, we want to use two-tailed wilcoxon rank sum test to test the significance.

H0: There is no difference between the pervalence differences of "after" period and that of "before" period.

HA: There is a difference between the pervalence differences of "after" period and that of "before" period.

```
# Calculate difference between each two years
ayvd <- data.frame(cbind(Q1_3$year,Q1_3$val))
colnames(ayvd) <- c("year","val")
```

```
ayvs2 <- split(ayvd,list(Q1_3$age))

for (k in 1:length(ayvs2)){
  for(cnt2 in 2:length(ayvs2[[k]]$year)){
  ayvs2[[k]][cnt2,"difference"] <-      ayvs2[[k]][cnt2,"val"]-ayvs2[[k]][(cnt2-1),"val"]
   }
}
# preform statistical tests
all_age_groups <- unique(Q1_3$age)
result <- data.frame()
for (i in all_age_groups){
  tmp <-
    wilcox.test(ayvs2[[i]]$difference[12:30],ayvs2[[i]]$difference[2:11])
  # Since nonparametric test tests medians between two groups,
  # we use medians to calculate the effect size
  ave <-
    median(ayvs2[[i]]$difference[12:30])-median(ayvs2[[i]]$difference[2:11])
  result <- rbind(result,c(i,tmp$p.value,ave))
}
names(result) <- c("age","p-value","effect size")
kable(result)
```

age	p-value	effect size
25 to 29	0.0770768461922885	0.00101540733885935
30 to 34	3.99400699250774e-07	0.00105993839470447
35 to 39	3.99400699250774e-07	0.000640406091123822
40 to 44	0.0090736849357539	0.000288887014208612
45 to 49	0.000387718228797689	0.000243001611018941
50 to 54	9.98501748126935e-08	0.000444739649946436
55 to 59	1.19820209775232e-06	0.000458942062569179
60 to 64	2.68596970246146e-05	0.000307529055261295
65 to 69	0.00112690907293606	0.000182938855314026

The result shows all groups except 25-29 group have p-values below 0.05. We can reject the null hypothesis and conclude that between the prevalence differences of "after" period and that of "before" period. And the effect size ("after"-"before") is larger than 0. Therefore we can confirm the increased opioid prescription can speed up the increase of the prevalence of opioid use disorders in all groups except 25-29 group.

The second part of the question (What age group is the most affected?)

For the second part of the question, we want to do linear regression for each period separately and compare the difference between the slopes of "before" and "after" periods. The age group with the largest difference is the most affected one.

We will take all groups which have shown the significance in the first part of the question.

From Figure 1.5 and 1.6, we can see the there is linear relationship (not "U" shaped etc..) between year and prevalence in all age groups. Also visual inspection shows that there is no obvious outliers. We think it is appropriate to do linear regression. We want to set the threshold of r.squared to 0.7 (from ADS2 week 18 lecture, slide 21). r.squared is used to evaluate the goodness of fit of the linear regression.

```r
# create data frame to store the result
temp <- Q1_3[which(Q1_3$age!="25 to 29"),] # eliminate 25 to 29 group
sig_age_groups <- all_age_groups[which(all_age_groups!="25 to 29")]
slope <- data.frame(matrix(nrow = length(sig_age_groups),ncol = 4))
rownames(slope) <- sig_age_groups
colnames(slope) <- c("before2000","r.squared for before 2000","after2000","r.squared for after 2000")
# separate data of before and after period for linear regression
temp[,"IF"] <- ifelse(temp$year >= 2000,1,0)
ayvs1 <- split(temp,list(temp$age,temp$IF))
cnt <- 0
for(i in c(1,3)) {
  for(j in sig_age_groups){
  cnt <- cnt + 1
  # do linear regression, get slope and r.squared
  fit <- lm(val~year,ayvs1[[cnt]])
  slope[j,i] <- as.numeric(fit$coefficients[2])
  slope[j,i+1] <- summary(fit)$r.squared
    }
}
# calculate the difference between slopes
slope$difference <- slope$after2000-slope$before2000
# arrange by difference
slope <- arrange(slope,desc(difference))
kable(slope)
```

	before2000	r.squared for before 2000	after2000	r.squared for after 2000	difference
30 to 34	0.0011604	0.9990004	0.0026010	0.9635391	0.0014406
35 to 39	0.0007209	0.9977019	0.0018841	0.9489746	0.0011631
40 to 44	0.0006921	0.9942977	0.0013555	0.9362292	0.0006633
50 to 54	0.0003175	0.9804609	0.0009381	0.9722804	0.0006206
55 to 59	0.0001644	0.9373350	0.0007299	0.9779954	0.0005655
45 to 49	0.0005473	0.9914923	0.0010876	0.9455326	0.0005403
60 to 64	0.0000799	0.7901554	0.0004977	0.9700913	0.0004178
65 to 69	0.0000526	0.6443105	0.0003101	0.9479582	0.0002575

Since the values of r.squared for all age groups except 65-69 ("before" period) are higher the threshold we previously set, the linear regression can reflect the data points. So we can compare the difference of slope among all age groups except 65-69.

Since 30-34 age group has the largest difference, we can confirm that this age is the most affected.

Note that we cannot establish causation between the increased opioid prescription and the changes of prevalence here, since we only have the data from the region where there is an increased opioid prescription. If we want to establish causation, we need to have a control group where the doctors are more careful when prescribing opioid.

Part 2: Our own question

To find something interesting from the data, we would like to first plot all the data and see how it looks like.

```
data_opioid <- data[which(data$cause=="Opioid use disorders"),]
data_death <- data_opioid[which(data_opioid$measure=="Deaths"),]
data_preval <- data_opioid[which(data_opioid$measure=="Prevalence"),]
g <- ggplot(data_death,aes(x=year,y=val,color=age))
g <- g + geom_line()
g <- g + facet_grid(sex~location)
g <- g + labs(title = "The death of opioid use disorders",
              y = "Prevalence (%)",caption = "Figure 2.1")
g <- g + theme(strip.text.x = element_text(angle = 90))
g <- g + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
g
```

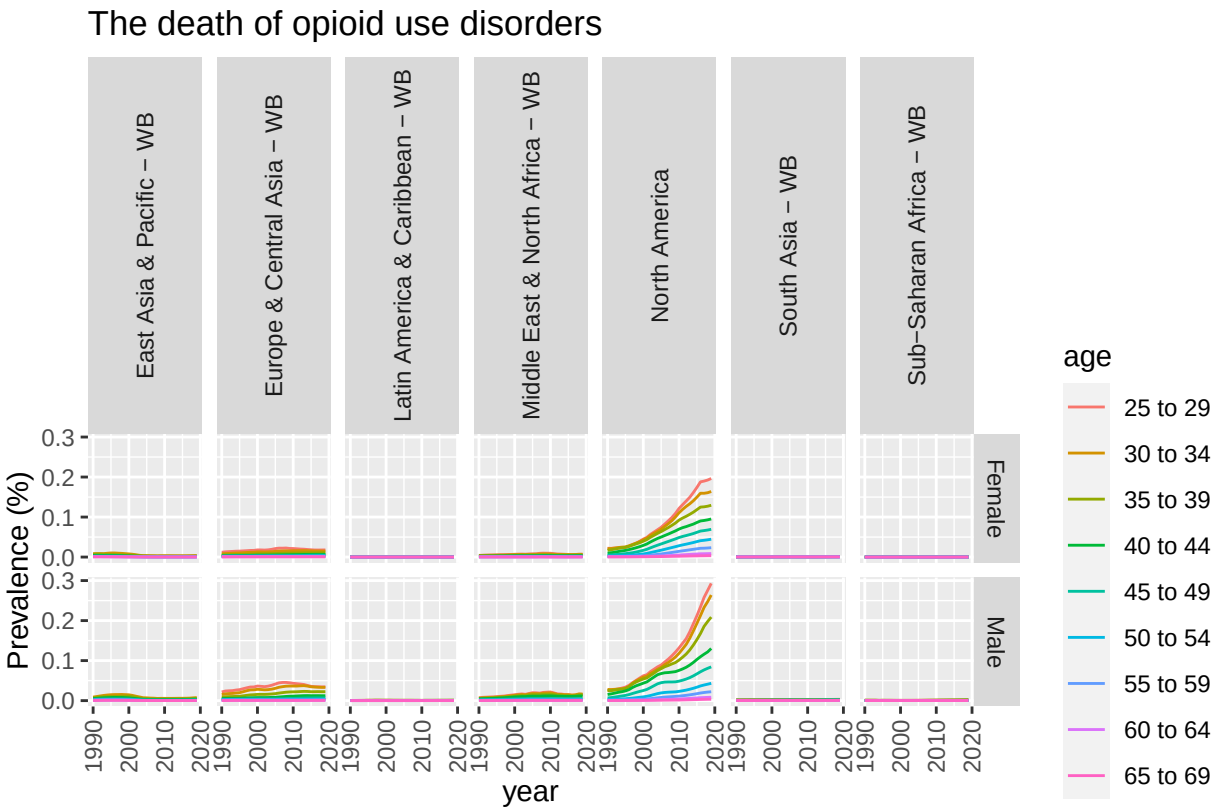

Figure 2.1

```
g <- ggplot(data_preval,aes(x=year,y=val,color=age))
g <- g + geom_line()
g <- g + facet_grid(sex~location)
g <- g + labs(title = "The prevalence of opioid use disorders",
              y = "Prevalence (%)",caption = "Figure 2.2")
g <- g + theme(strip.text.x = element_text(angle = 90))
g <- g + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
g
```

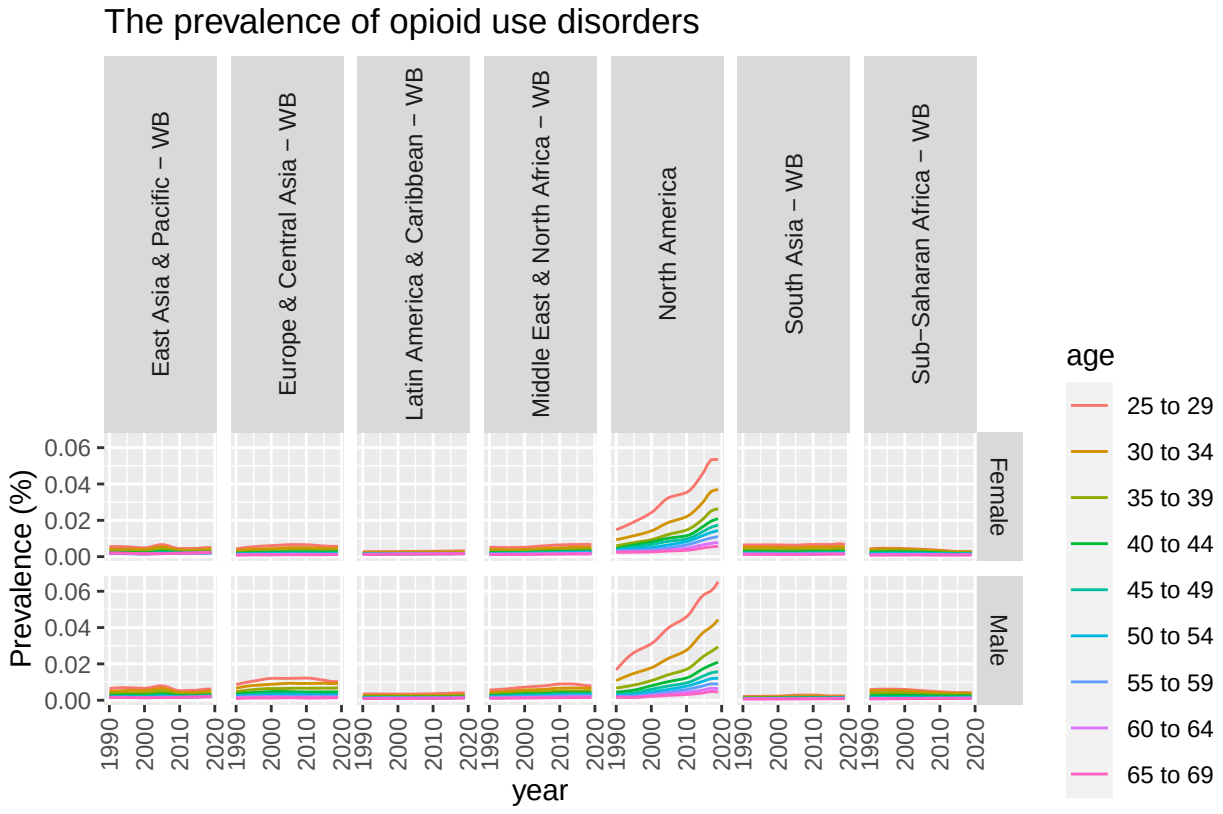

Figure 2.2

```
data_alcohol <- data[which(data$cause=="Alcohol use disorders"),]
data_death <- data_alcohol[which(data_alcohol$measure=="Deaths"),]
data_preval <- data_alcohol[which(data_alcohol$measure=="Prevalence"),]
g <- ggplot(data_death,aes(x=year,y=val,color=age))
g <- g + geom_line()
g <- g + facet_grid(sex~location)
g <- g + labs(title = "The death of alcohol use disorders",
              y = "Prevalence (%)",caption = "Figure 2.3")
g <- g + theme(strip.text.x = element_text(angle = 90))
g <- g + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
g
```

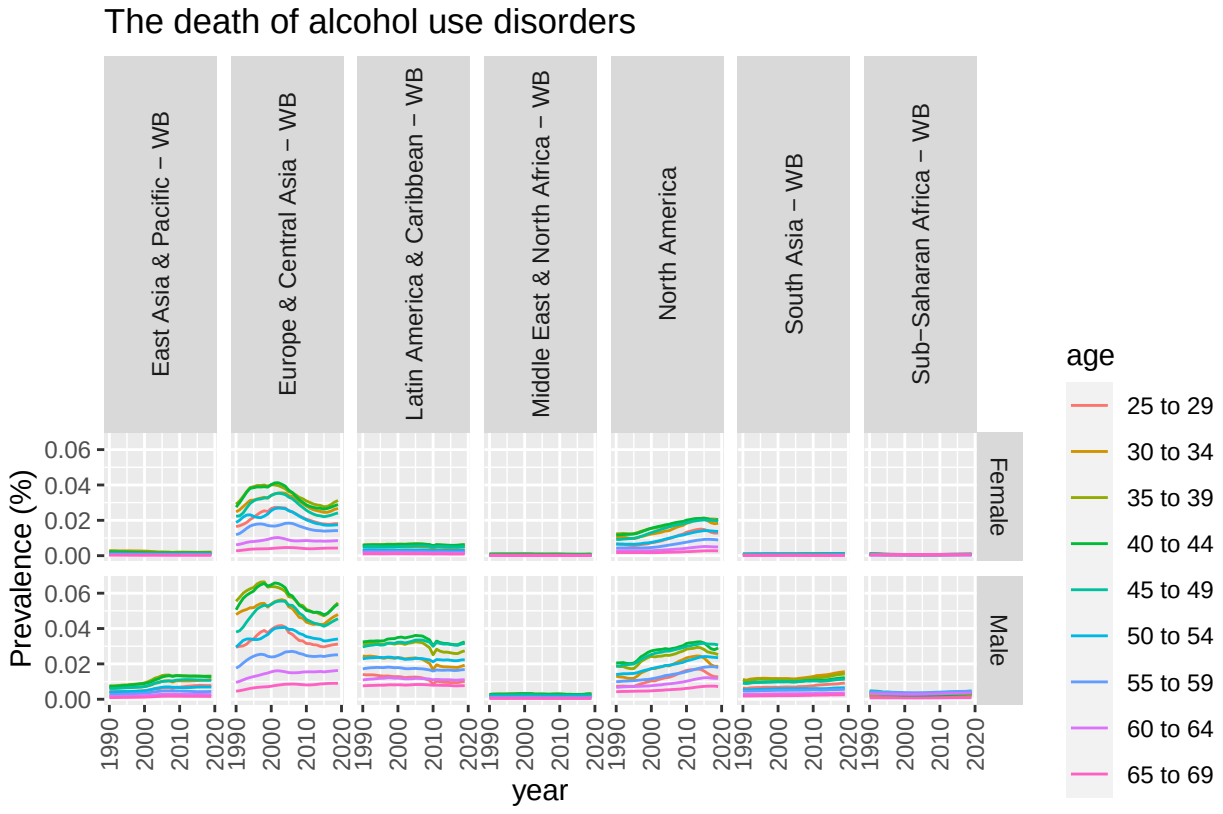

Figure 2.3

```
g <- ggplot(data_preval,aes(x=year,y=val,color=age))
g <- g + geom_line()
g <- g + facet_grid(sex~location)
g <- g + labs(title = "The prevalence of alcohol use disorders",
              y = "Prevalence (%)",caption = "Figure 2.4")
g <- g + theme(strip.text.x = element_text(angle = 90))
g <- g + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
g
```

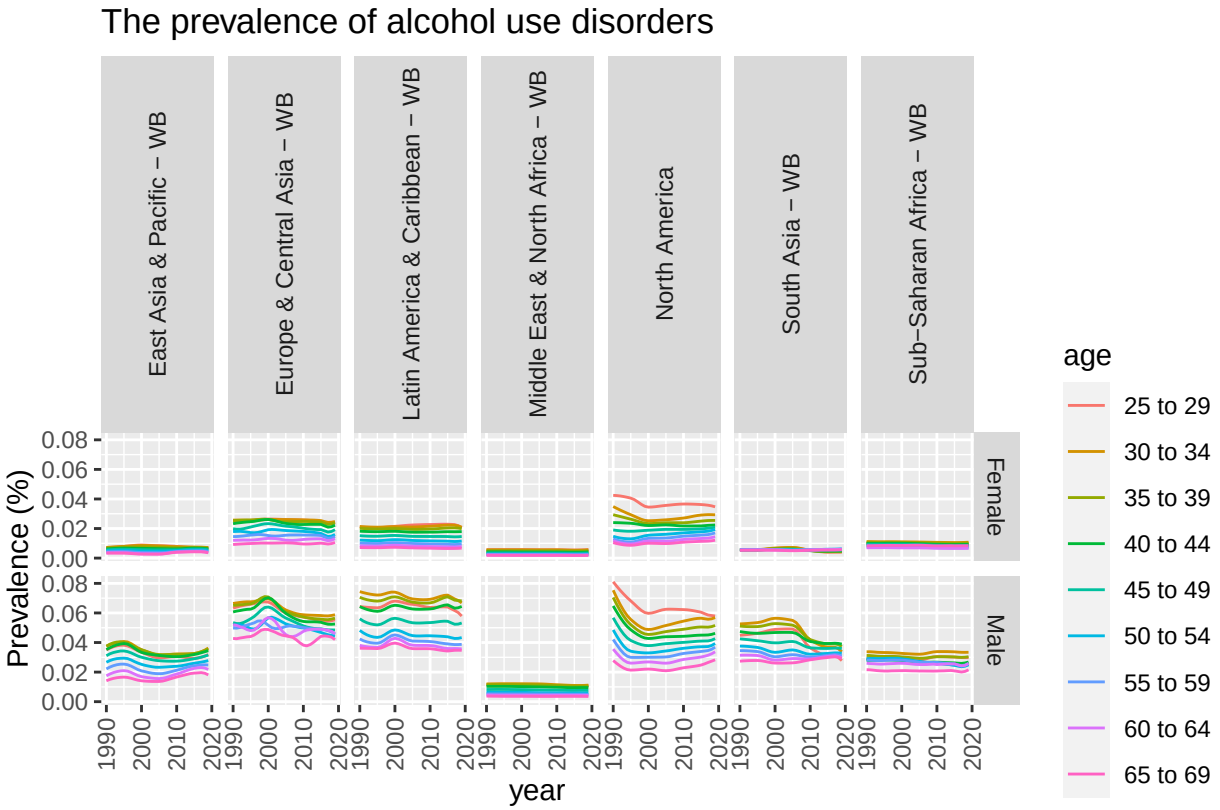

Figure 2.4

Question

From figure 2.4, we noticed generally male have larger alcohol use disorders than female in 1990-2019 in all locations and age groups.

We want to confirm our observation using statistical method.

Why we are interested in this question?

If male have larger alcohol use disorders than female worldwide, it is possible that the government or WHO need to have more strict policy on alcohol use on male.

Solution

We extract relevant data.

```
Q2 <- data[which(data$cause=="Alcohol use disorders" & data$measure=="Prevalence"),]
```

We want to use bootstrapping to confirm our hypothesis (since we can always use bootstrapping).

We plan to use case resampling bootstrapping. We pool the prevalence value of male and female together (total 60 data points). We randomly sample 30 data point from it with replacement twice to get two data sets and then calculate the median difference of the two sets. We do this for 10000 times and get a null distribution. We calculate the percentage of the data points in null distribution that is more extreme than the actual median difference between male and female. The percentage is the p-value.

H0: There is no difference between the prevalence of male and female in 1990-2019 in a certain age group in a certain location.

HA: There is a difference between the prevalence of male and female in 1990-2019 in a certain age group in a certain location.

```
all_location <- unique(data$location)
all_age <- unique(data$age)
total_year <- length(unique(data$year))
res <- data.frame()
for (i in all_location){
  temp <- Q2[which(Q2$location==i),]
  mean <- c()
  for(j in all_age){
    temp_M <- temp[which(temp$age==j & temp$sex=="Male"),]
    temp_F <- temp[which(temp$age==j & temp$sex=="Female"),]
    median <- median(temp_M$val)-median(temp_F$val)
    temp_all <- c(temp_M$val,temp_F$val)
    mean_null <- c()
    for (h in 1:10000) {
      sample1_null <- sample(temp_all,total_year,replace = TRUE)
      sample2_null <- sample(temp_all,total_year,replace = TRUE)
      mean_null <- c((median(sample1_null)-median(sample2_null)),mean_null)
    }
    mean <- c(mean,mean(mean_null>median))
  }
  res <- rbind(res,c(i,mean))
}
names(res) <- c("location",all_age)
kable(res)
```

location	25 to 29	30 to 34	35 to 39	40 to 44	45 to 49	50 to 54	55 to 59	60 to 64	65 to 69
East Asia & Pacific - WB	0	1e-04	2e-04	6e-04	4e-04	5e-04	1e-04	3e-04	1e-04
North America	0	0	0	0	3e-04	0	0	0	0
Middle East & North Africa - WB	1e-04	1e-04	2e-04	2e-04	0	2e-04	3e-04	0	0
Europe & Central Asia - WB	3e-04	8e-04	5e-04	0	2e-04	0	1e-04	0	0
Latin America & Caribbean - WB	0	1e-04	0	1e-04	0	0	0	0	0
Sub-Saharan Africa - WB	4e-04	0	0	4e-04	8e-04	3e-04	3e-04	0	0
South Asia - WB	1e-04	0	0	0	2e-04	4e-04	1e-04	1e-04	3e-04

Although multiple testing increase the false positive rate, the p-values are all quite low. We can still conclude there is a difference between the prevalence of male and female in 1990-2019 worldwide. Visual inspection (Figure 2.4) shows that male have larger prevalence than female. We can conclude that generally male have larger alcohol use disorders than female in 1990-2019 in all locations and age groups.

Also, we plot the data with confidence interval.

```
all_location <- unique(data$location)
cnt <- 4
for (i in all_location){
  cnt <- cnt + 1
  temp <- Q2[which(Q2$location==i),]
  g <- ggplot(temp,aes(year,val,color=sex))
  g <- g + geom_errorbar(aes(ymin=lower, ymax=upper))
  g <- g + geom_line(size=1)
  g <- g + geom_point(size=0.5)
  g <- g + labs(title=paste0("Prevalence of alcohol use disorders ",i),y="Prevalence (%)",caption = pas
  g <- g + facet_wrap(~age)
  g <- g + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=.5))
  plot(g)
}
```

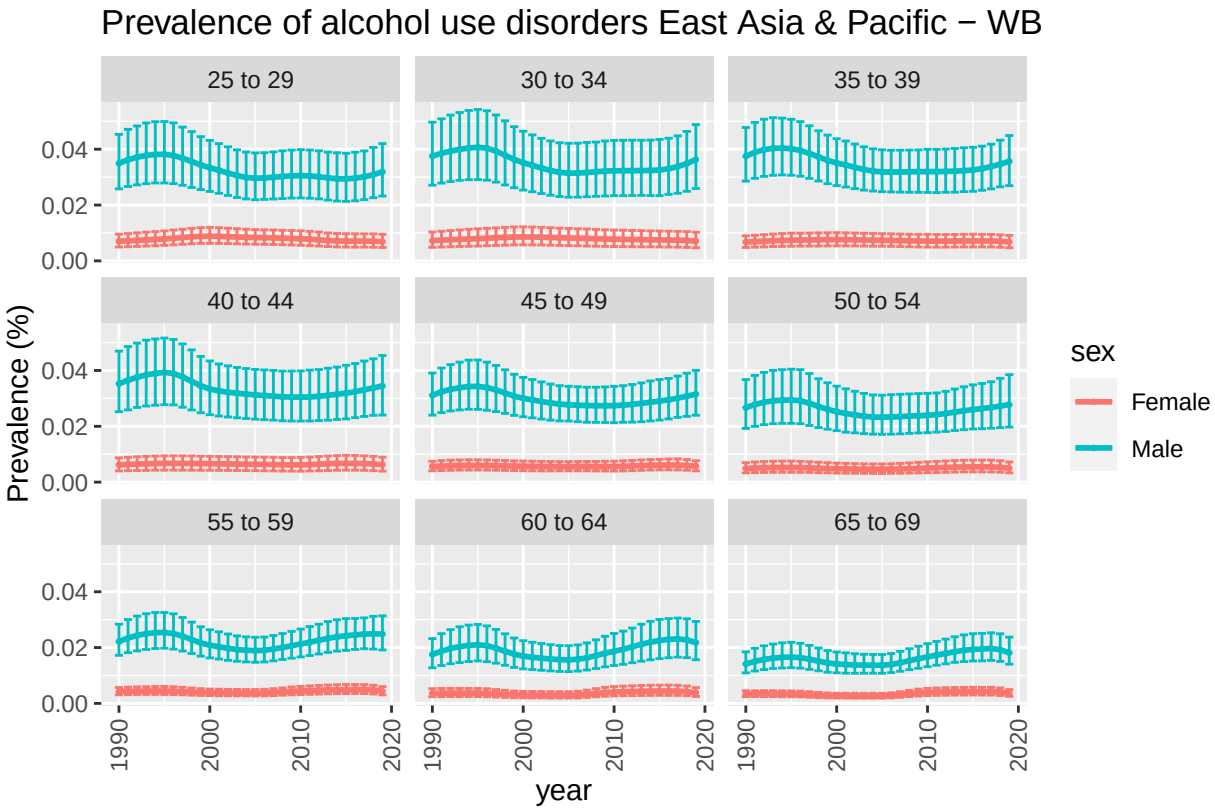

Figure 2.5

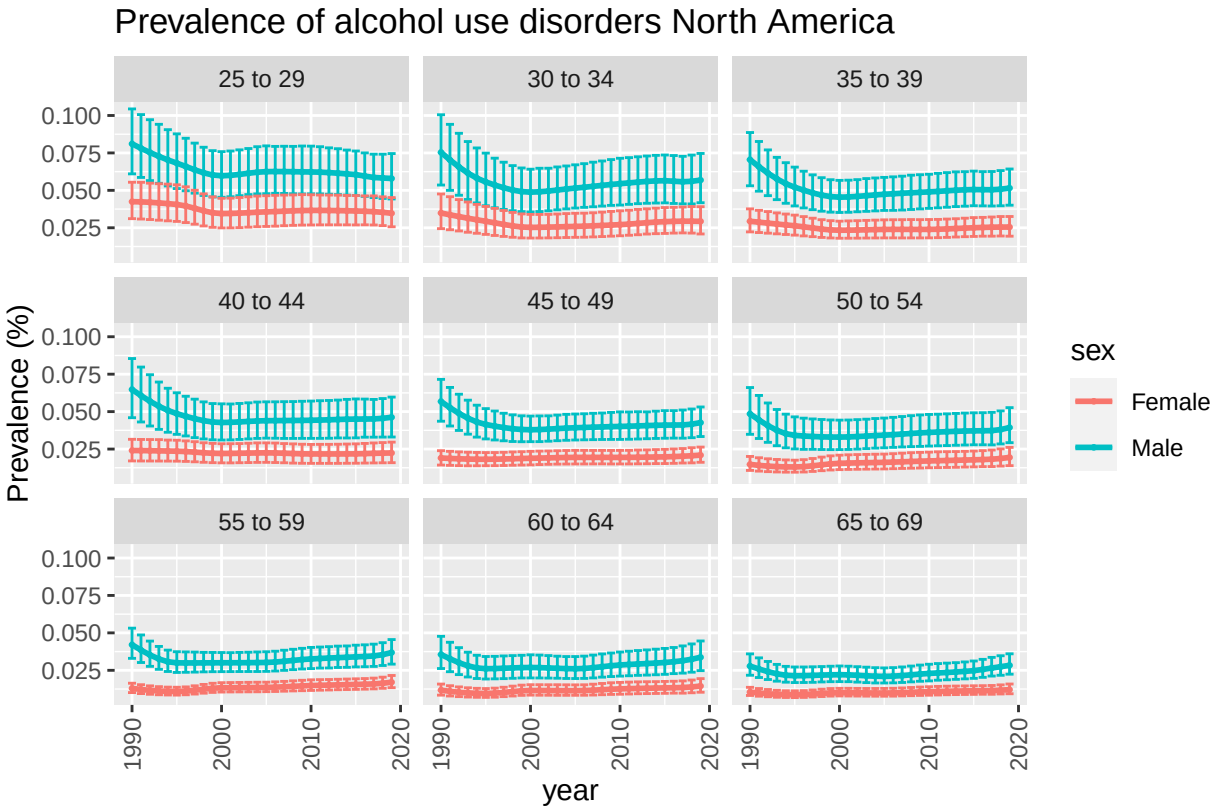
Prevalence of alcohol use disorders North America
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69
Prevalence (%)
0.100
0.075
0.050
0.025
1990
2000
2010
2020
year
sex
Female
Male

Figure 2.6

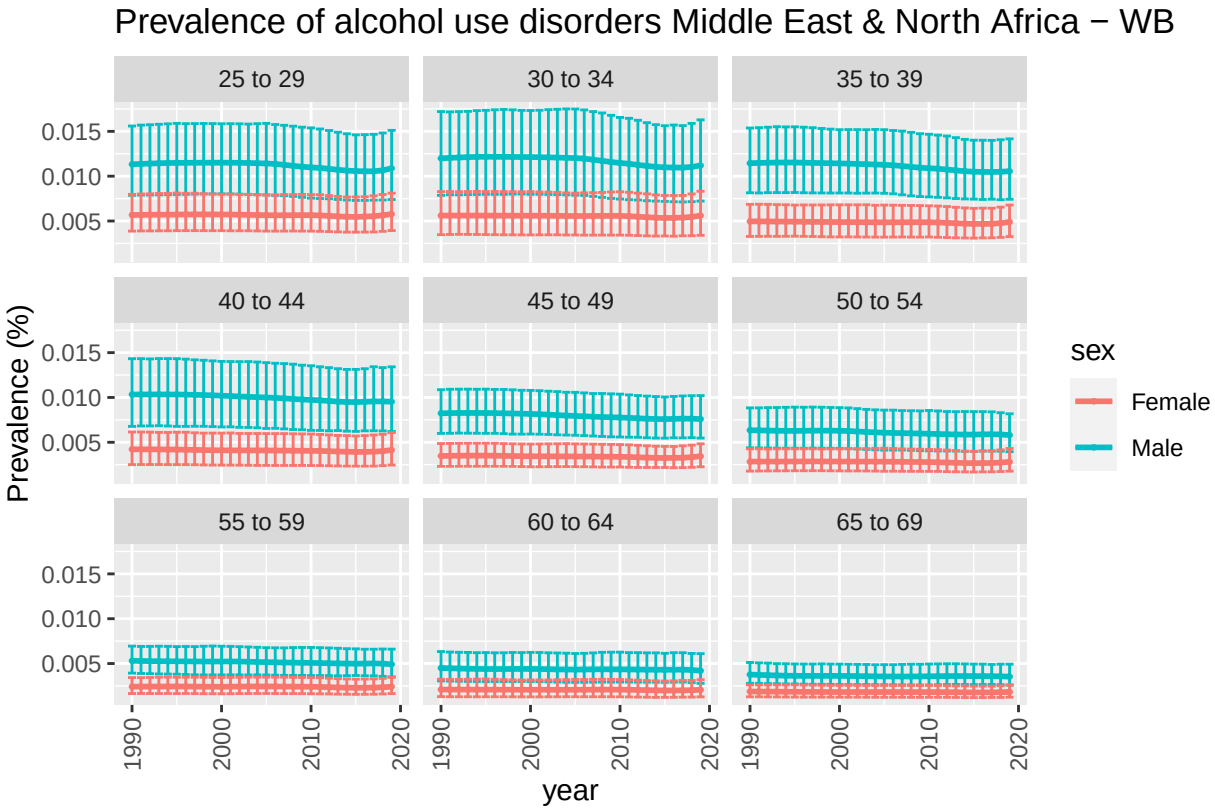
Prevalence of alcohol use disorders Middle East & North Africa − WB
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69
Prevalence (%)
0.015
0.010
0.005
1990
2000
2010
2020
year
sex
Female
Male

Figure 2.7

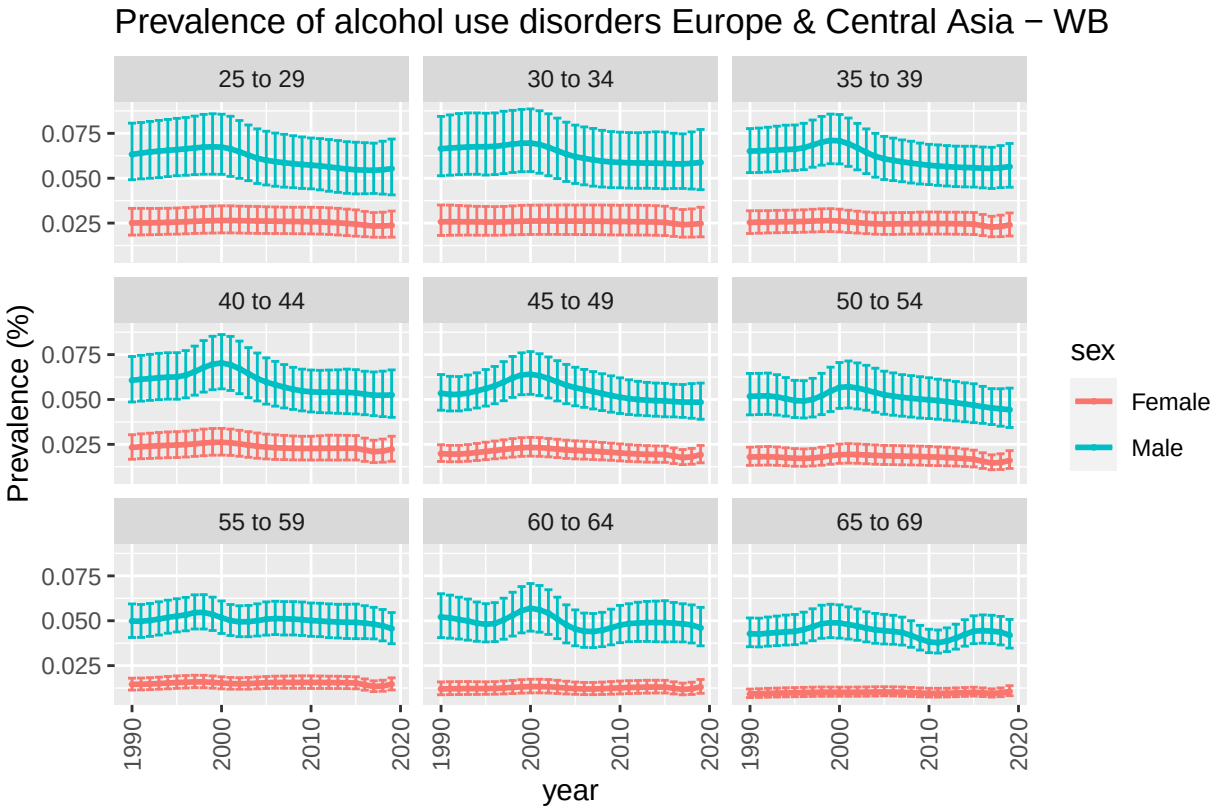
Prevalence of alcohol use disorders Europe & Central Asia − WB
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69
Prevalence (%)
0.075
0.050
0.025
1990
2000
2010
2020
year
sex
Female
Male

Figure 2.8

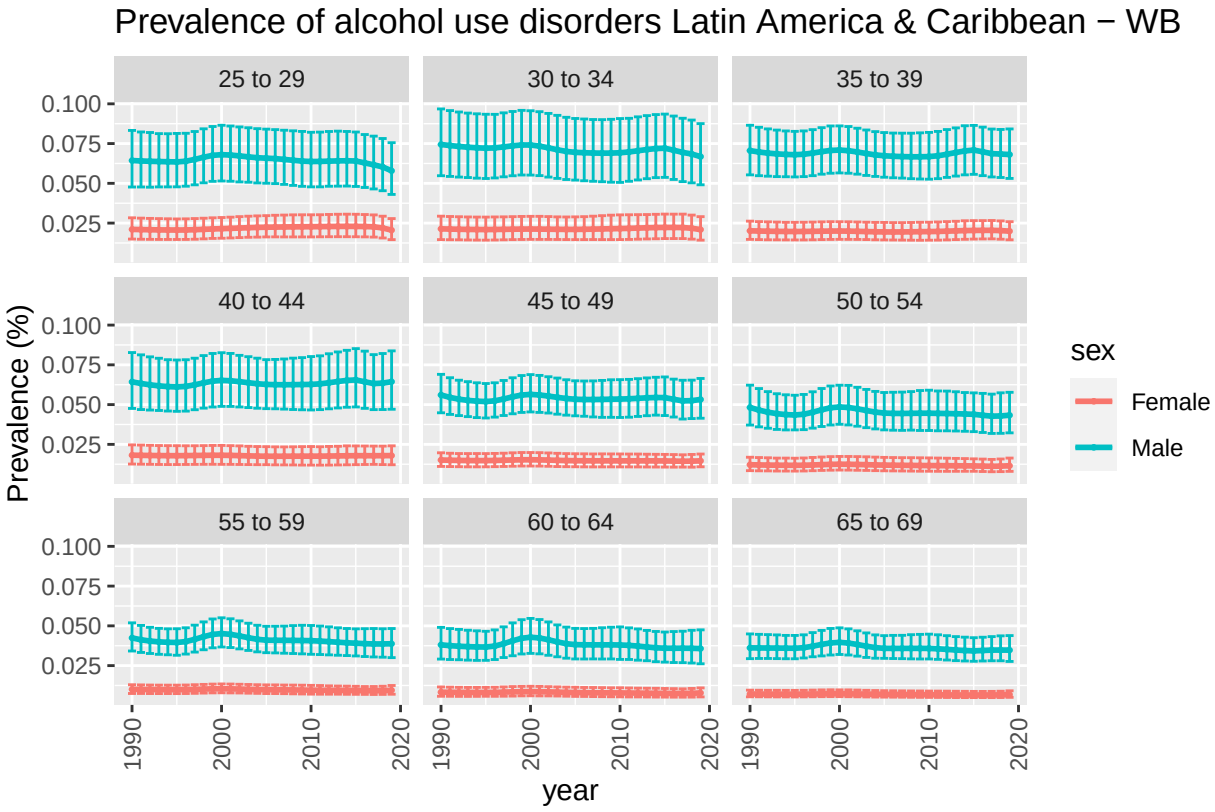
Prevalence of alcohol use disorders Latin America & Caribbean − WB
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69
Prevalence (%)
0.100
0.075
0.050
0.025
1990
2000
2010
2020
year
sex
Female
Male

Figure 2.9

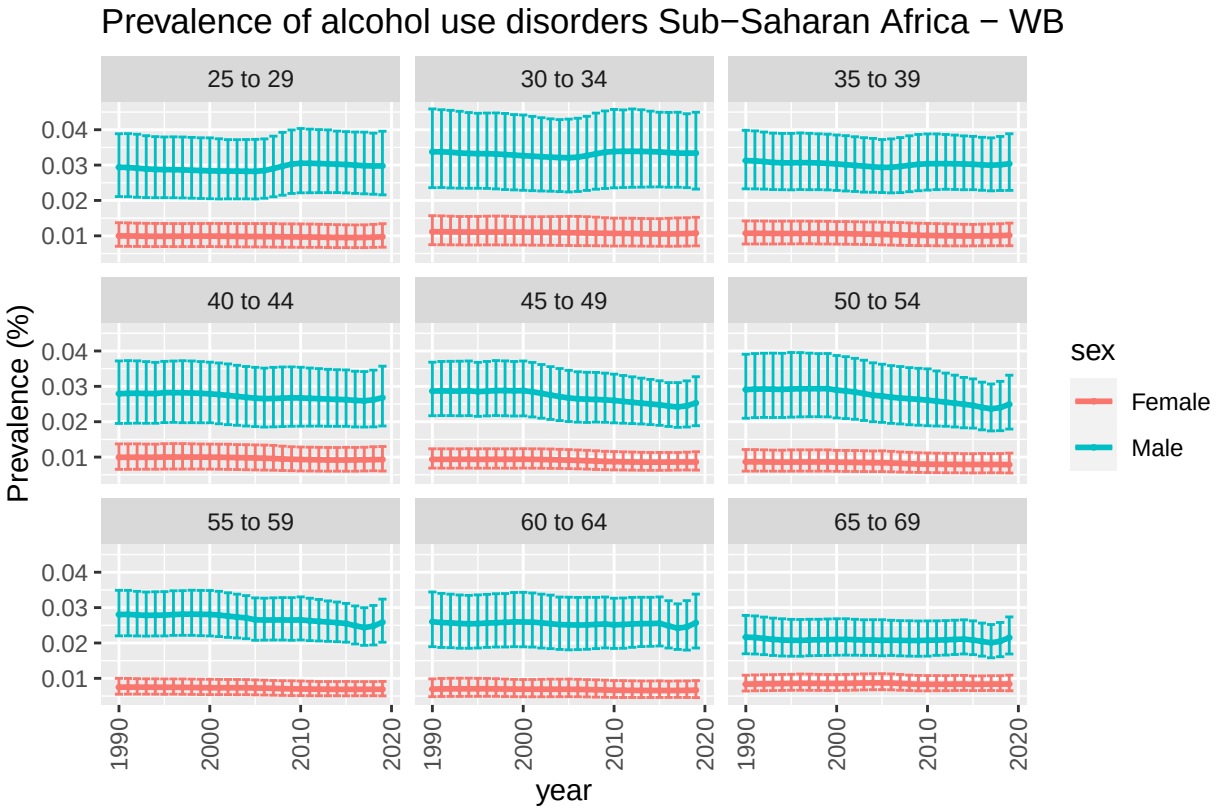
Prevalence of alcohol use disorders Sub−Saharan Africa − WB
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69
Prevalence (%)
0.04
0.03
0.02
0.01
1990
2000
2010
2020
year
sex
Female
Male

Figure 2.10

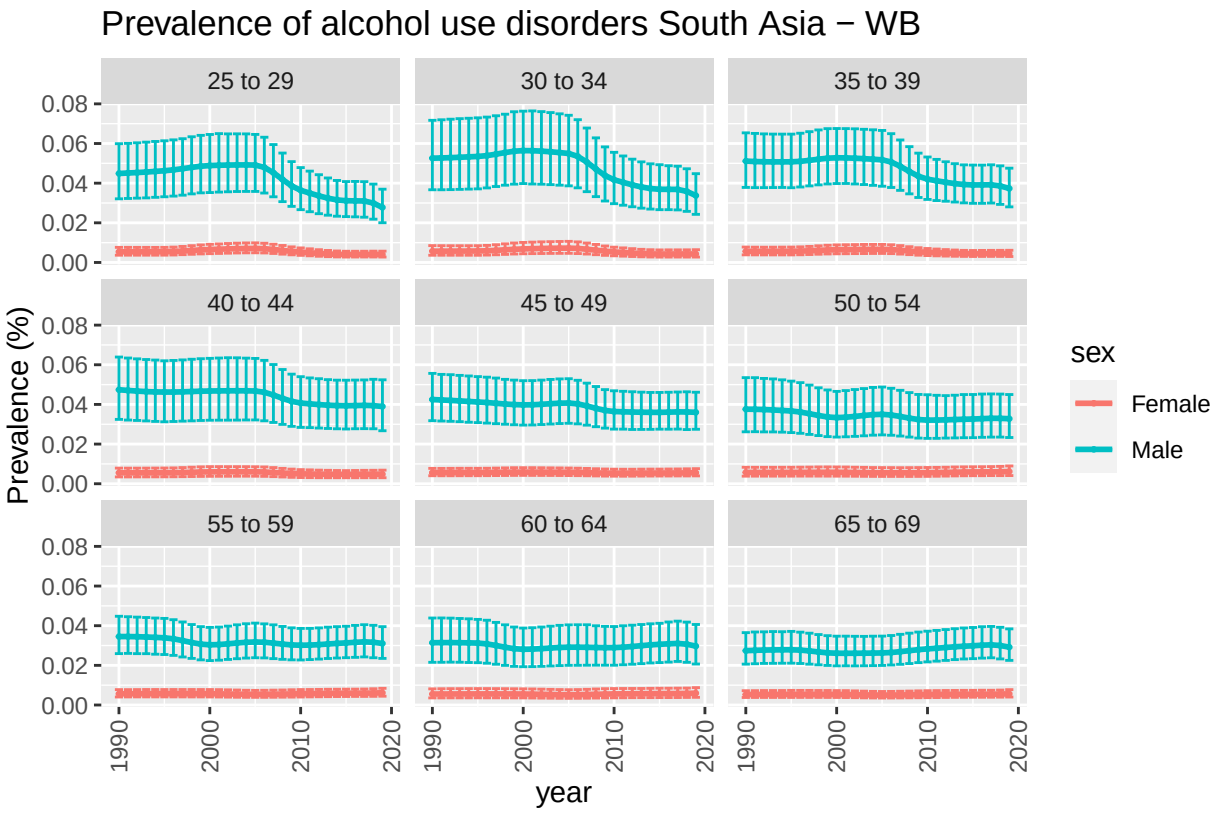

Figure 2.11

We found that there is a little overlapping (Figure 2.6, 2.7) between men and women but majority has no overlap, which further confirm our conclusion.