ADS_Group_Project

9080; 9064; 9009; 9003; 9026; 9091

2021/4/25

Brief Introduction

ADS Group Project aims to explore the prevalence and death rate of diseases related to the use of alcohol and opioids. To know more information about our project, you can visit our project on GitHub (https://github.com/ADS-Group-Exercise2/ADS2-Group-Exercise2). We restored our supplementary files there, including:

- our project R Markdown file ("ADS2_GroupExercise2.Rmd")
- three pictures ("country.png","percentage.png", "world.png") and the table ("substance_use.csv") used in the R Markdown file
- PDF file generated using MiKTeX

Import Data and libraries

```
dt <- read.csv("substance_use.csv")
library(dplyr)
library(tidyr)
library(ggplot2)
```

Part 1: Exploring the data

Q1.

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

We use the filter function to filter the information about alcohol-related deaths among men aged 40-44 in the year 2019 and then sort the highest rate in the world.

```
data1<-filter(dt,sex=="Male"&age=="40 to 44"
              &cause=="Alcohol use disorders"&measure=="Deaths"&year=="2019")
data1[which.max(data1$val), ]$location
```

```
## [1] Europe & Central Asia - WB
## 7 Levels: East Asia & Pacific - WB ... Sub-Saharan Africa - WB
```

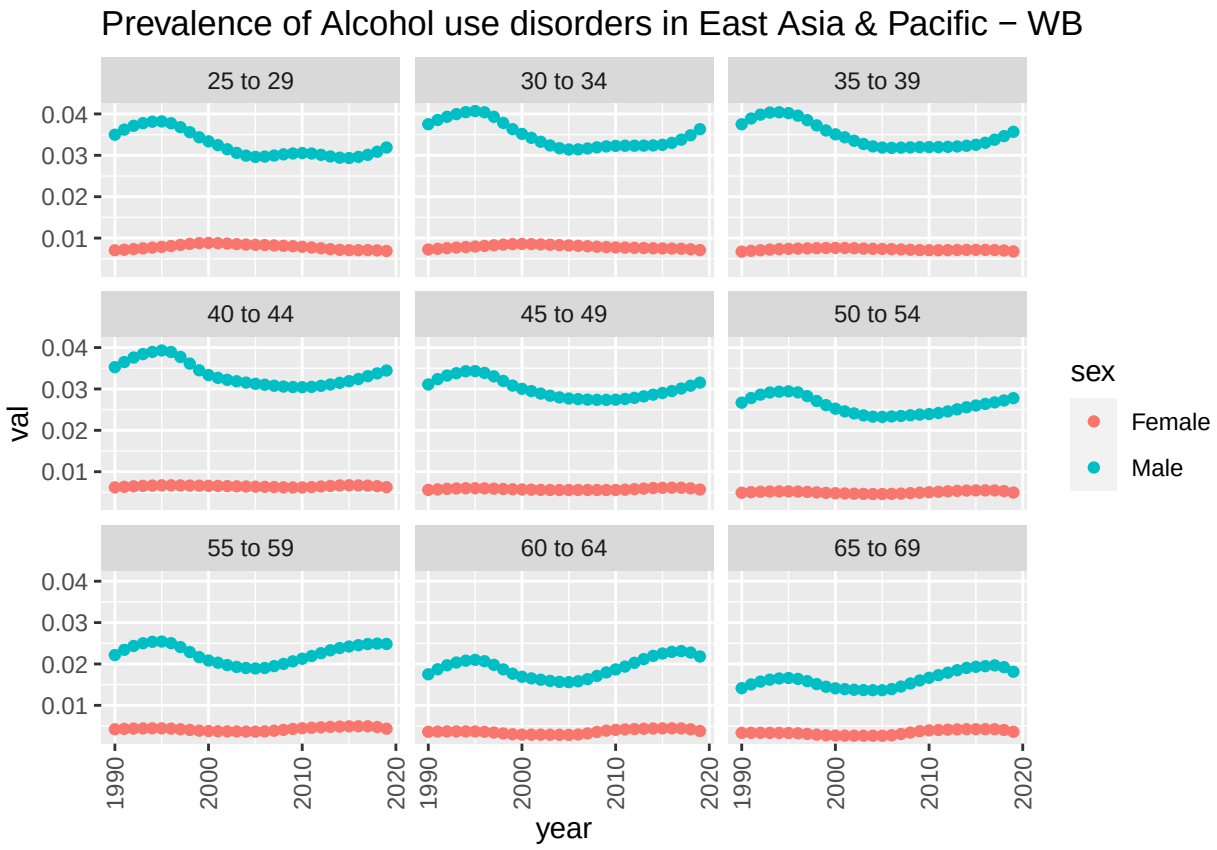
Here we find that Europe & Central Asia - WB region has the highest rate of alcohol-related deaths among men aged 40-44.

Q2.

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? Is there a difference between men and women?

The first step is to select the data we want for further analysis. Here we use index with which function to do that.

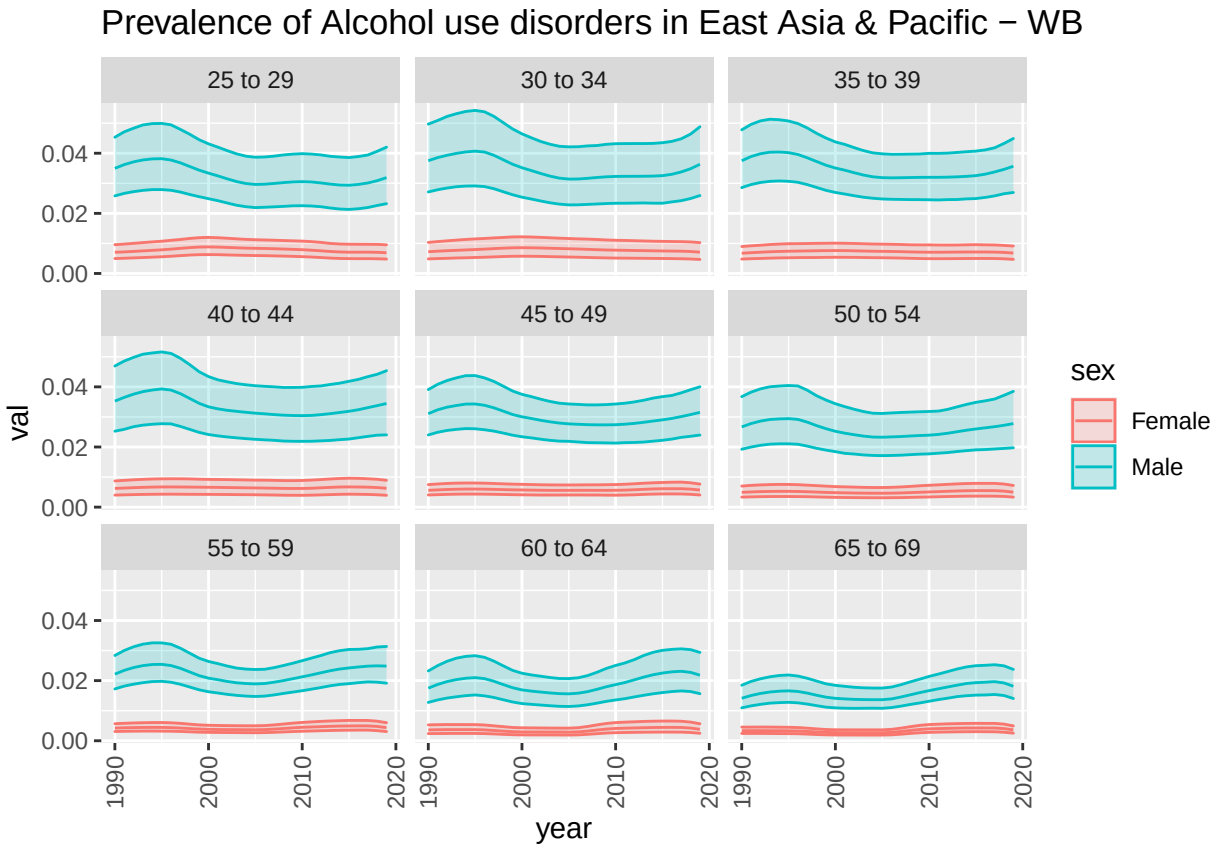
```
data2<-dt[which(dt$cause=="Alcohol use disorders"&
                  dt$location=="East Asia & Pacific - WB"&dt$measure=="Prevalence"),]
#Plot the changes over the years in different age groups and sex
ggplot(data2,aes(x=year,y=val))+
  geom_point(aes(color=sex))+
  ggtitle("Prevalence of Alcohol use disorders in East Asia & Pacific - WB")+
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=.5))+facet_wrap(~age)
```



From the plot, it can be generally seen that in female and all age groups, the prevalence of alcohol use disorders in East Asia & Pacific did not change significantly, and was relatively steady from 1990 to 2020, while the prevalence for male first increased (1990~1995) and then decreased (1995~2000), and kept steady until a slight increase was seen after 2015. Specifically, the prevalence of alcohol use disorders in 25 to 29, 30 to 34, 35 to 39, 40 to 44, and 45 to 49 were similar, which was around 0.01 for females and between 0.03 and 0.04 for males. The prevalence of the 50 to 54 age group (Female:~ 0.005; Male: 0.025~0.03) was slightly lower than the previous 5 age groups but was slightly higher than the other 3 age groups (55 to 59, 60 to 64, 65 to 69). The prevalence of female was similar in the age group 55 to 59, 60 to 64, and 65 to 69 (i.e.<0.005), while the prevalence of male was similar in the age group 55 to 59 and 60 to 64 (0.02~0.025), but was slightly lower in the age group 65 to 69 (0.015~0.02). Notably, in all age groups, the prevalence of males was markedly higher than that of females, and we may consider that there was a difference between men and women referring to the prevalence of alcohol use disorders in the East Asia & Pacific region.

Additionally, we noticed that upper and lower bounds to the confidence interval around the value were provided, so we also plot the data with the confidence interval.

```
ggplot(data2, aes(x = year, color = sex, fill = sex)) +
  geom_ribbon(aes(ymin = lower, ymax = upper), alpha = 0.2) +
  geom_line(aes(y = val))+
  facet_wrap(~age)+
  ggtitle("Prevalence of Alcohol use disorders in East Asia & Pacific - WB")+
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=.5))
```



It seems that we can get a similar conclusion from this plot. Besides, it seems that the margin of error of 55 to 59, 60 to 64, and 65 to 69 are smaller than the other age groups.

Q3.

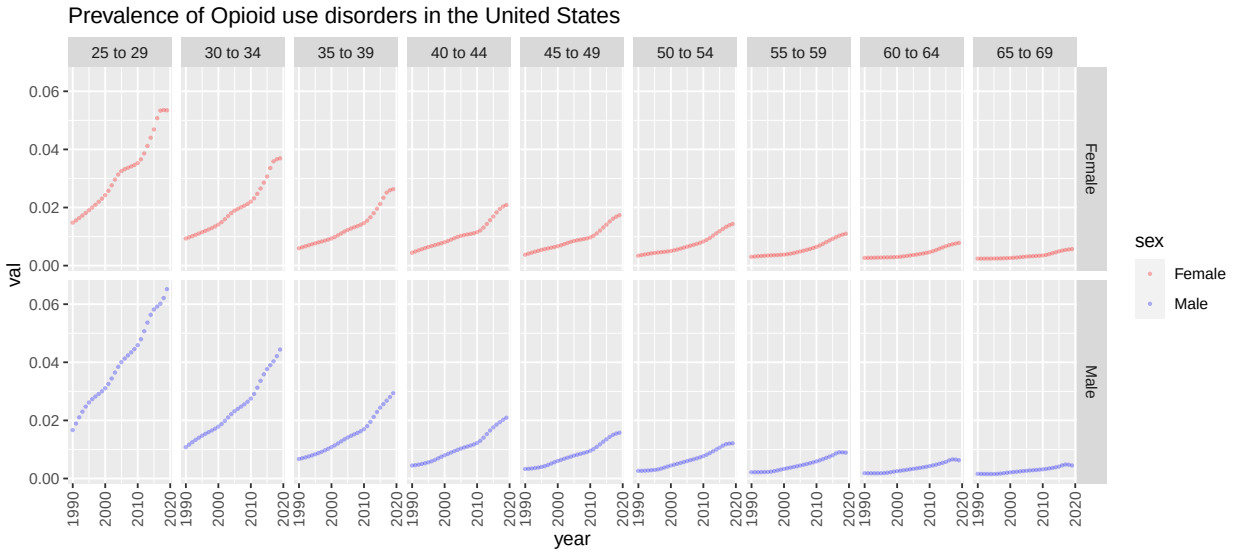
In the United States, there is talk of an "Opioid epidemic". Part of the problem is that since the late 1990s, doctors have increasingly been prescribing pain killers which can be highly addictive. Looking at the data from the United States, can you confirm an increase in the prevalence of diseases related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

Firstly, we use the filter function to filter the information about opioid-related prevalence in the United States. To investigate this question, much information is unrelated and we just focus on "sex", "age", "year", "val" columns.

```
opioid<-filter(dt, measure == "Prevalence", location == "North America",
               cause == "Opioid use disorders")
opioid<-opioid[c('sex','age','year','val')]
```

Then perform the age categorized visualization.

```
ggplot(opioid)+
  geom_point(aes(year,val,color=sex),alpha=0.3,size=0.5)+facet_grid(sex~age)+
  ggtitle("Prevalence of Opioid use disorders in the United States")+
  scale_colour_manual(values=c("red","blue"))+
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=.5))
```



From the plot, it can be seen that in all age groups, the prevalence of opioid use disorders in the United States kept increasing from 1990 to 2020. The prevalences among different age groups are not similar, people with lower age tend to be affected more by opioid use.

To identify the age group with the highest increasing rate for the prevalence of opioid use disorders, we use linear regression as the method since the figure above shows that there tends to be a linear correlation between the year and the prevalence value.

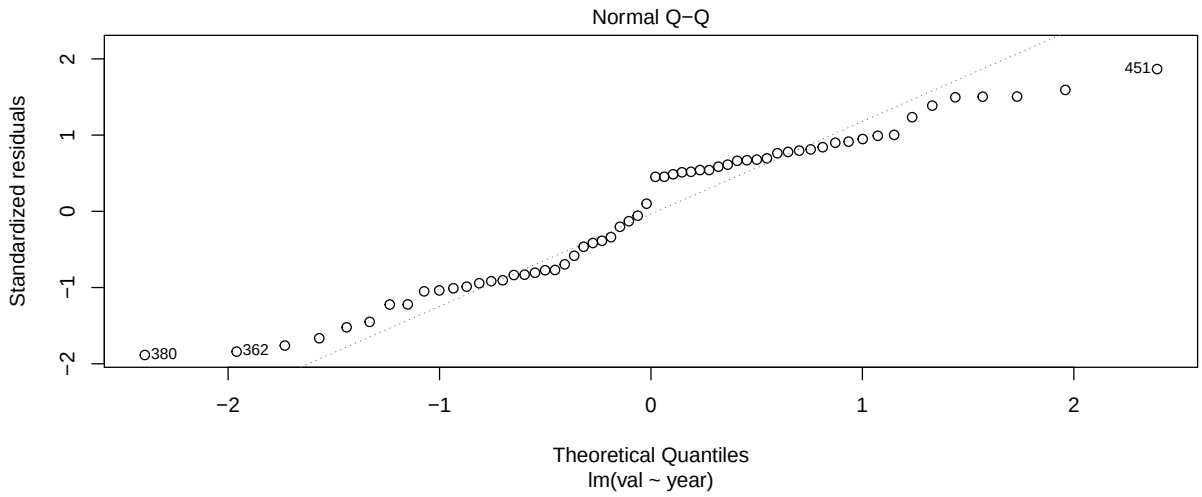
To apply the linear regression, we need to check the assumptions for the linear regression for every age group. There are three assumptions for the linear regression:

1. Independent random sampling
2. Normality of residuals
3. Equality of variances

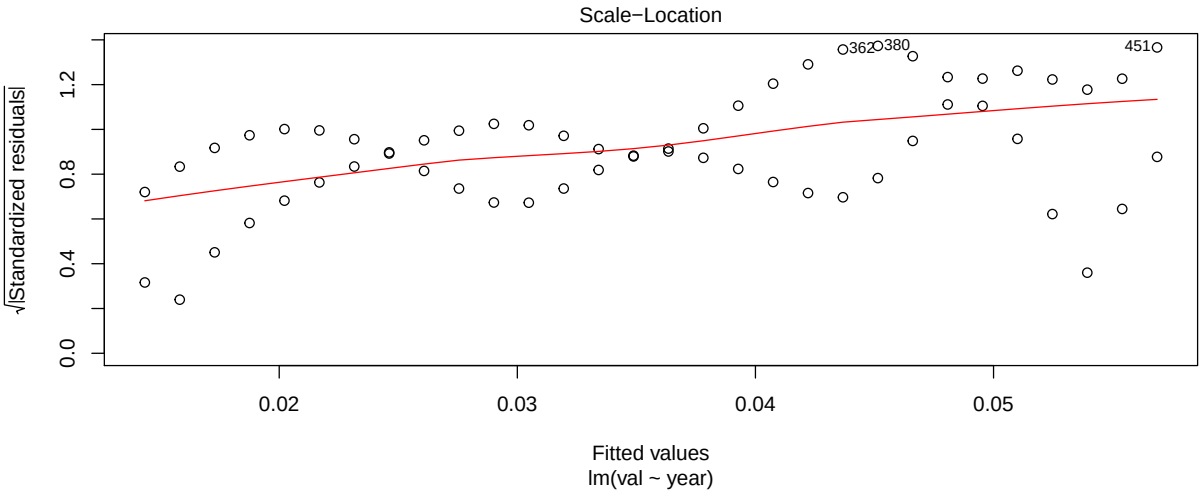
- For the first assumption, as the data provided can not be used to assess that whether the data is independent random sampling, thus we need to believe that this assumption is true based on the description of the experiment and survey about opioid use disorders.
- For the other two assumptions, the figures that show the equality of variances and normality of residuals for each age group are plotted below.

```
vec<-unique(opioid$age)
for(i in 1:9){
  opioid_tmp<-opioid[which(opioid$age==vec[i]),]
  opioid_l<-lm(val~year,data=opioid_tmp)
  plot(opioid_l,2,main = paste("Age group:",vec[i],"-(Normality of residuals)"))
  plot(opioid_l,3,main = paste("Age group:",vec[i],"-(Equality of variances)"))
}
```

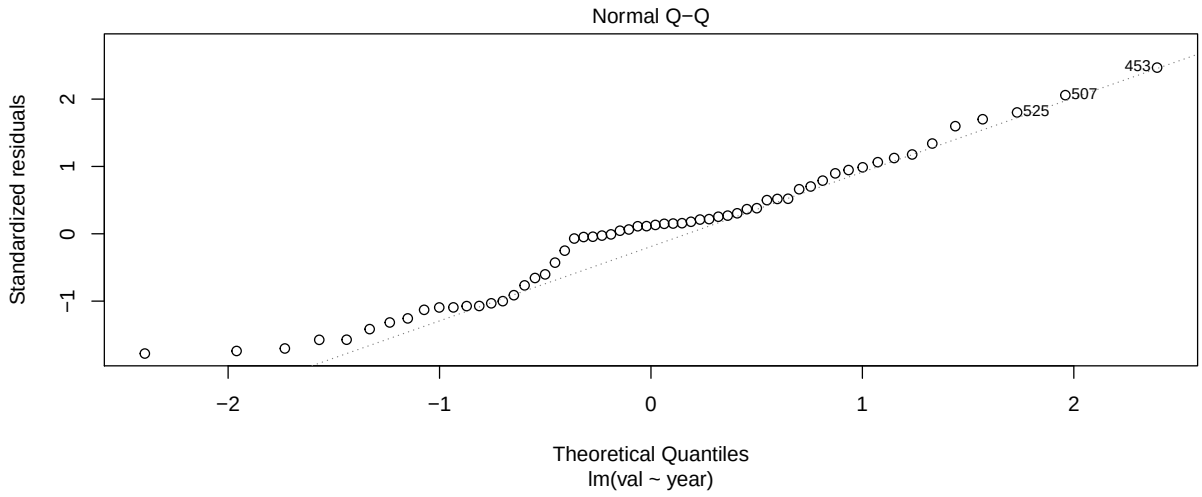
Age group: 25 to 29 −(Normality of residuals)



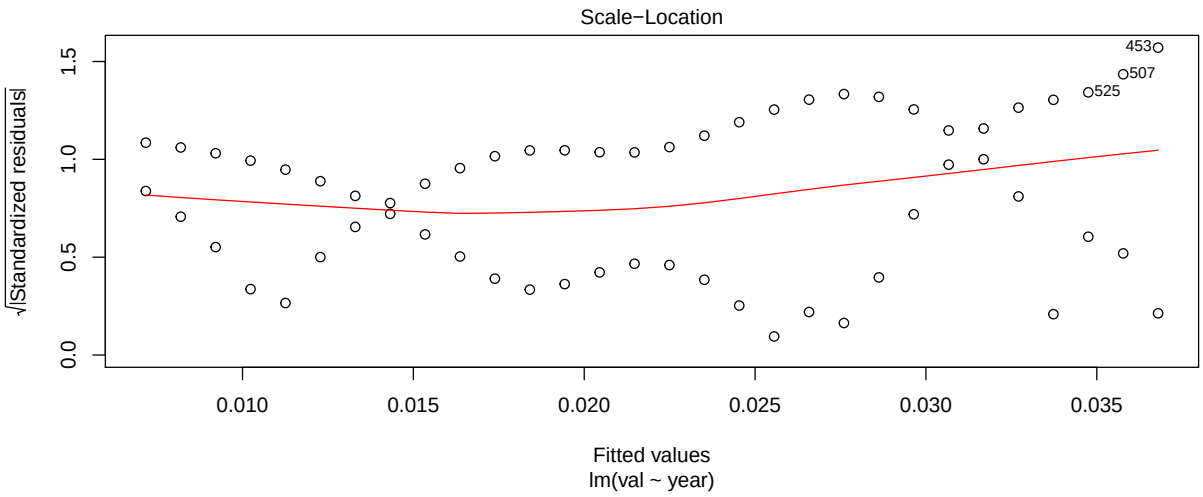
Age group: 25 to 29 −(Equality of variances)



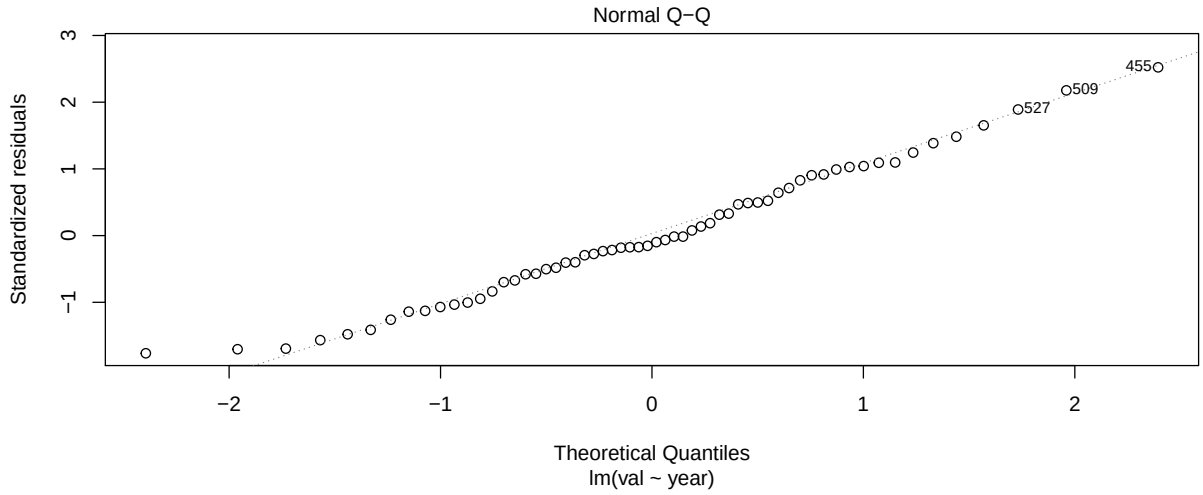
Age group: 30 to 34 −(Normality of residuals)



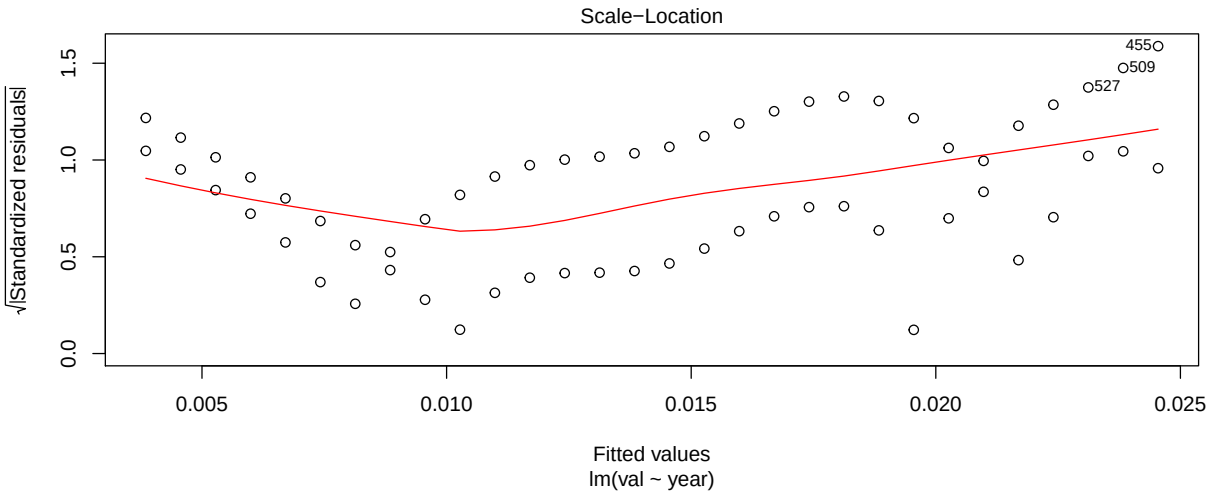
Age group: 30 to 34 −(Equality of variances)



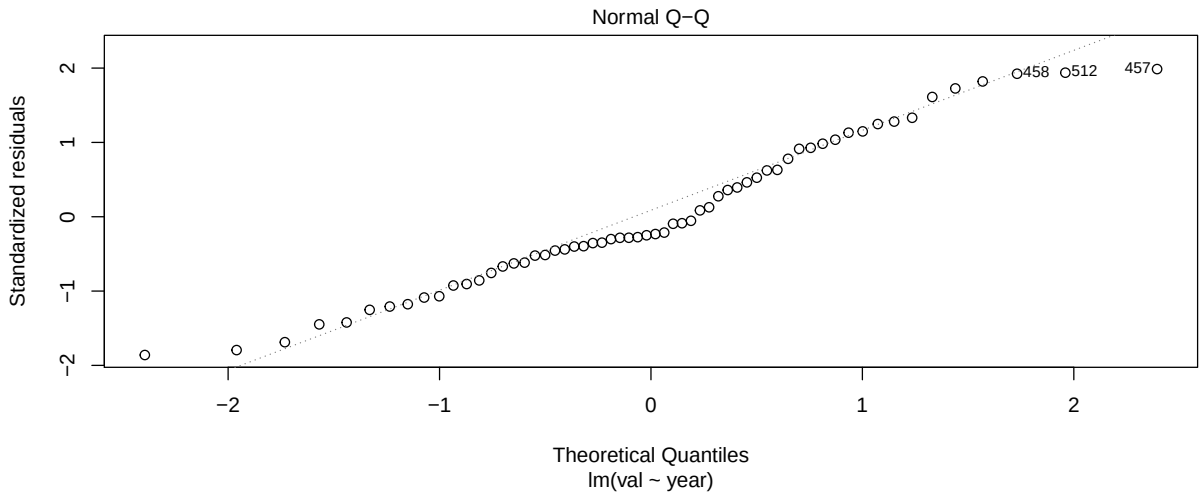
Age group: 35 to 39 −(Normality of residuals)



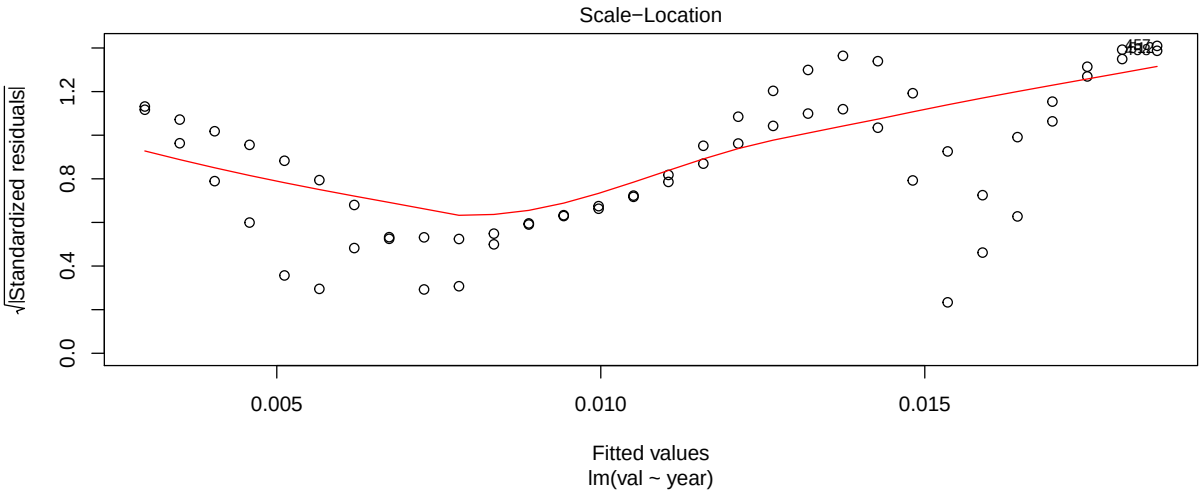
Age group: 35 to 39 −(Equality of variances)



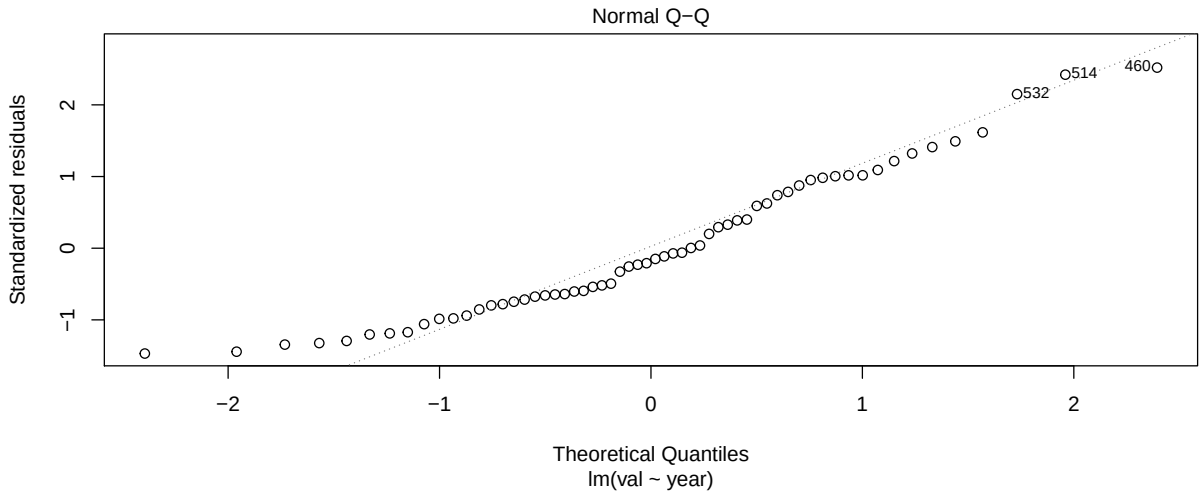
Age group: 40 to 44 −(Normality of residuals)



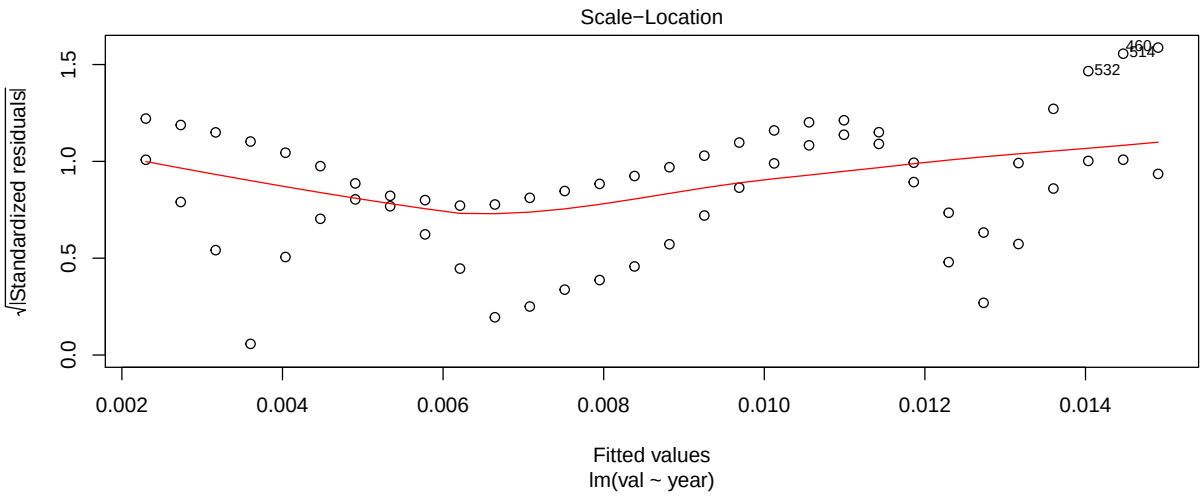
Age group: 40 to 44 −(Equality of variances)



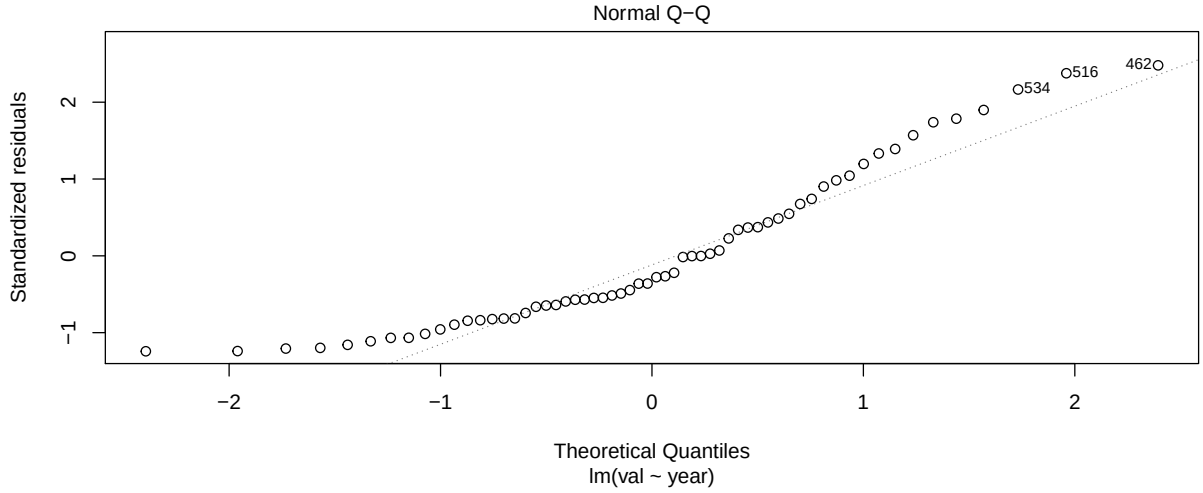
Age group: 45 to 49 −(Normality of residuals)



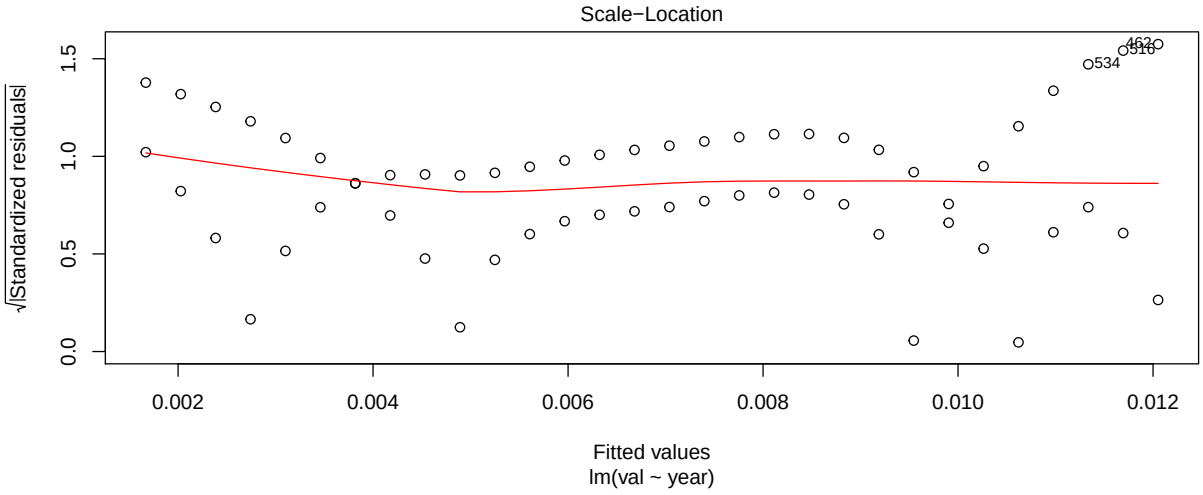
Age group: 45 to 49 −(Equality of variances)

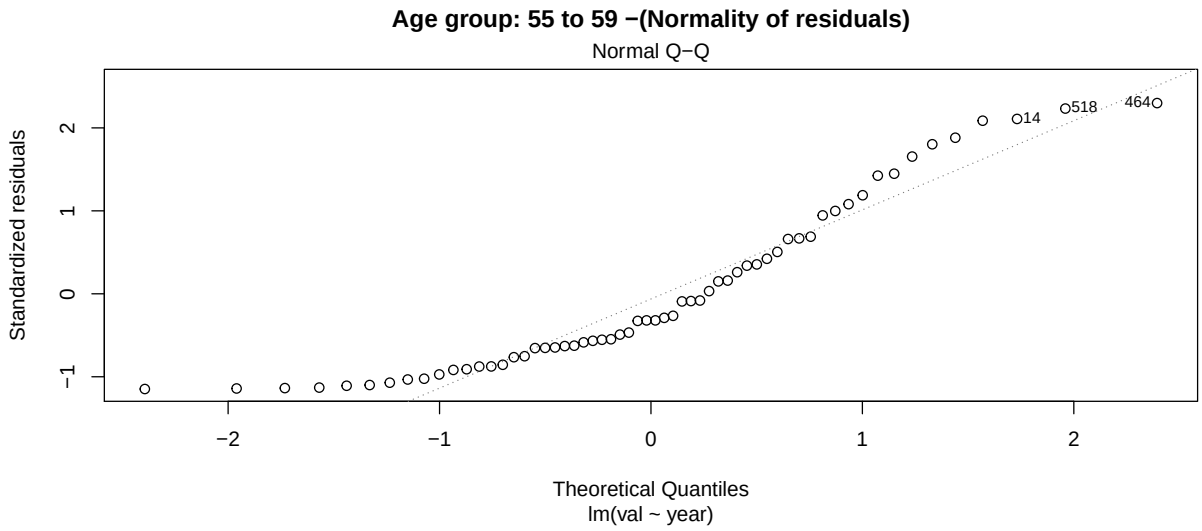


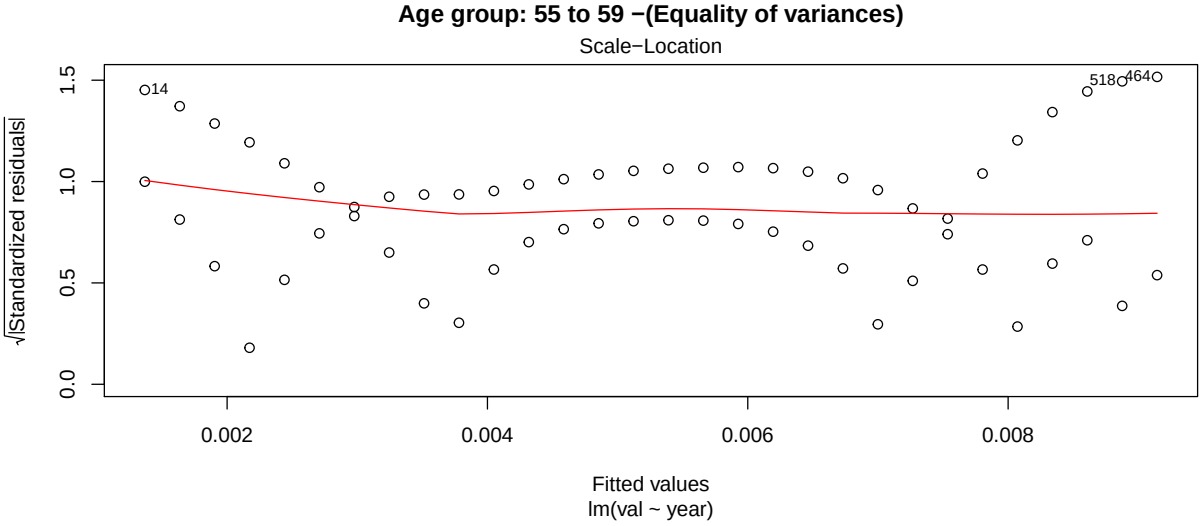
Age group: 50 to 54 −(Normality of residuals)

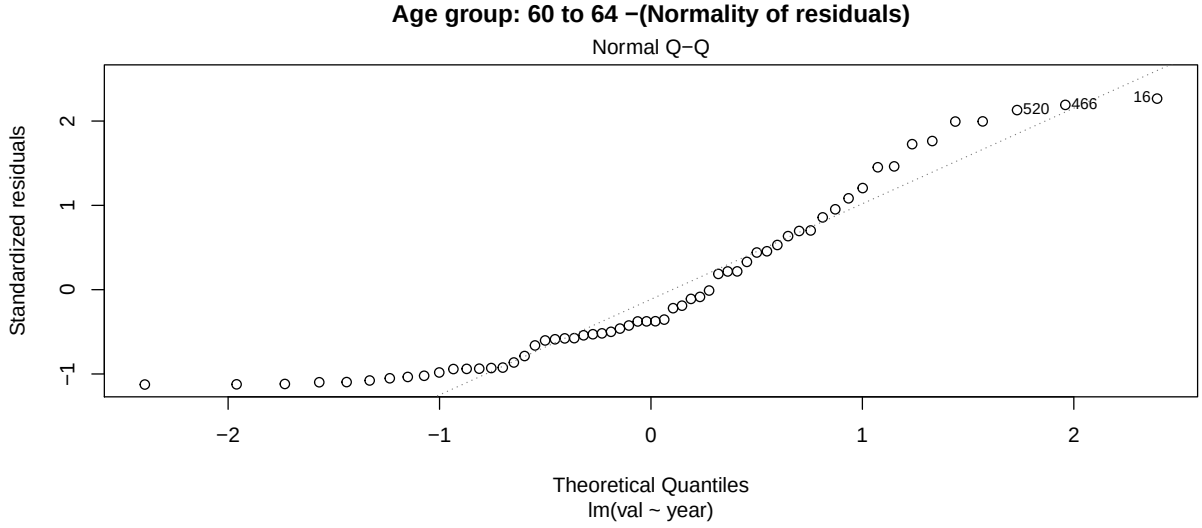


Age group: 50 to 54 −(Equality of variances)

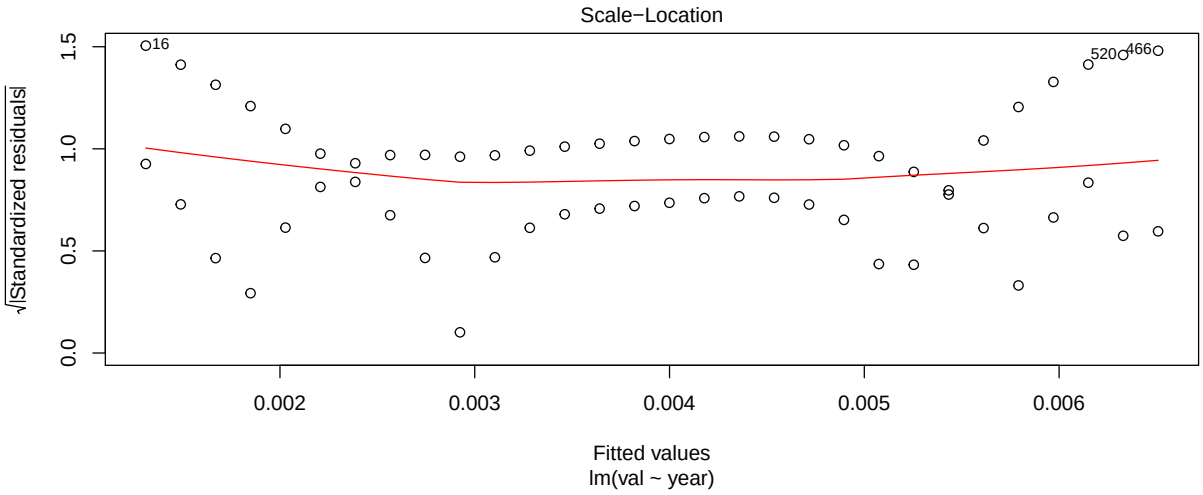



Age group: 55 to 59 −(Normality of residuals)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(val ~ year)
14
518
464

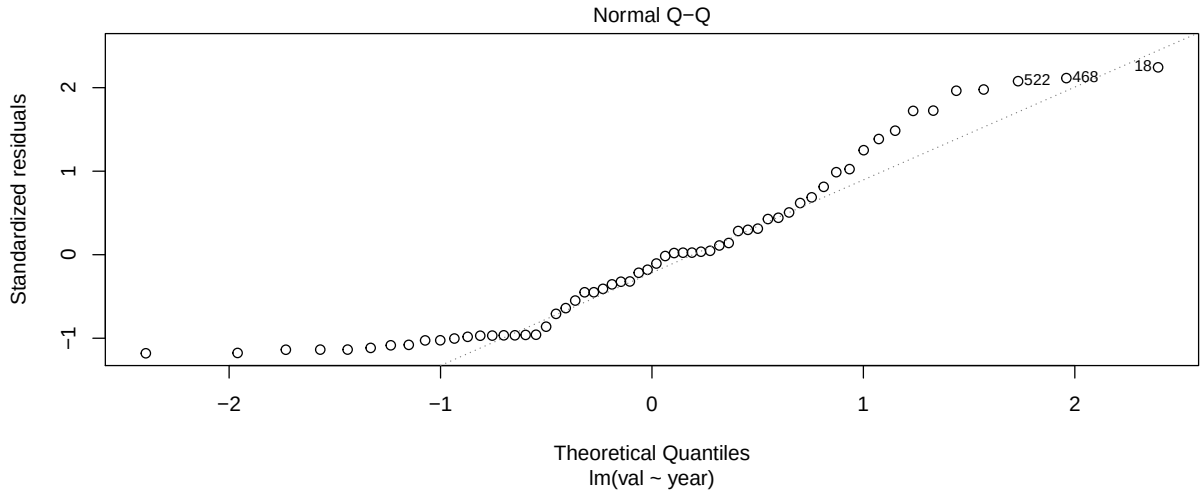

Age group: 55 to 59 −(Equality of variances)
Scale−Location
√|Standardized residuals|
Fitted values
lm(val ~ year)
14
518
464


Age group: 60 to 64 −(Normality of residuals)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(val ~ year)
520
466
16

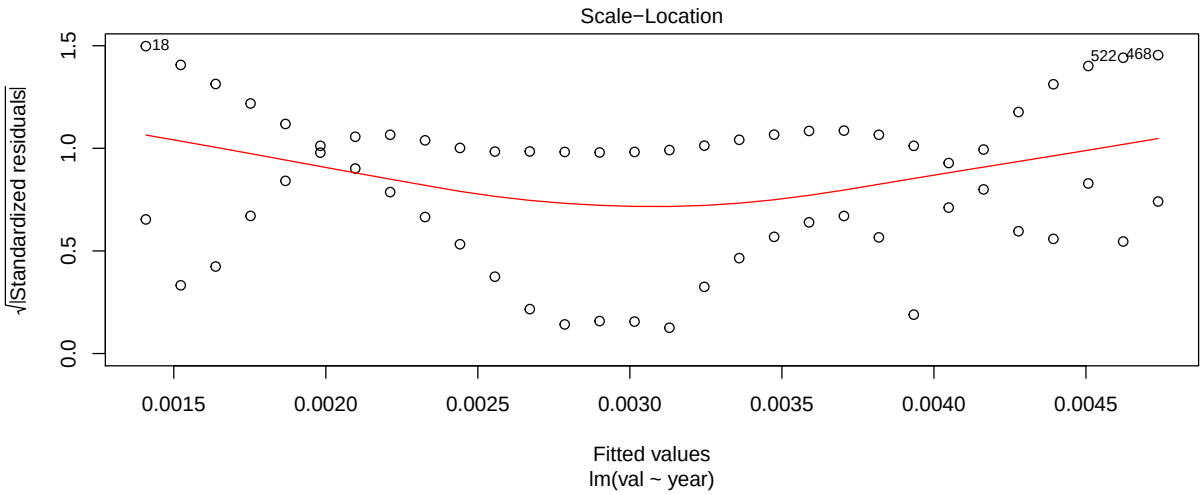
Age group: 60 to 64 −(Equality of variances)



Age group: 65 to 69 −(Normality of residuals)



Age group: 65 to 69 −(Equality of variances)

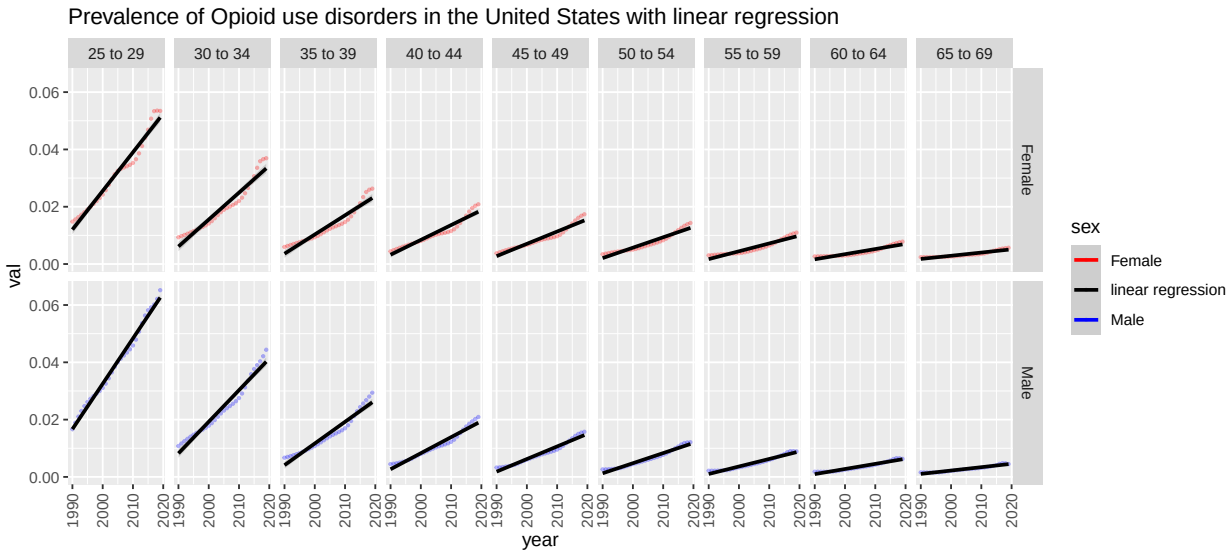


The result shows that the data of all age groups fulfill the requirement of normality of residuals and equality of variances, thus the assumptions for the linear regression are met.

After verifying the assumptions, we run the linear regression and collect the results.

- Plotting the figure with the regression line

```
ggplot(opioid)+
  geom_point(aes(year,val,color=sex),alpha=0.3,size=0.5)+
  facet_grid(sex~age)+
  stat_smooth(aes(year,val,color="linear regression"),method = lm)+
  scale_size_manual(values=0.1)+
  ggtitle("Prevalence of Opioid use disorders in the United States with linear regression")+
  scale_colour_manual(values=c("red","black","blue"))+
  theme(axis.text.x = element_text(angle=90, hjust=1, vjust=.5))
```



- Performing the linear regression

```
vec<-unique(opioid$age)
# Letter "n" at the end indicates that the variable is a list of the elements.
kn<-c() #The slope of the regression line representing the changing rate;
pn<-c() #The p-value from T-statistic to test the significance for the variable;
#R squared and adjusted R squared representing how well the regression was done;
Rn<-c(); Radjn<-c()
fpn<-c() #The p-value from F-statistics to test the significance for the whole equation.
for(i in 1:9){
  opioid_tmp<-opioid[which(opioid$age==vec[i]),]
  opioid_l<-lm(val~year,data=opioid_tmp)
  sy<-summary(opioid_l)
  k<-sy$coefficients[2,1]; p<-sy$coefficients[2,4]
  R<-sy$r.squared; Radj<-sy$adj.r.squared
  fp<-pf(sy$fstatistic[1L], sy$fstatistic[2L], sy$fstatistic[3L], lower.tail = FALSE)
  kn<-c(kn,k); pn<-c(pn,p); Rn<-c(Rn,R)
  Radjn<-c(Radjn,Radj); fpn<-c(fpn,fp)
}
res<-cbind(vec,kn,pn,Rn,Radjn,fpn)
```

After getting results, put them in a data frame and adjust the format and column name to make it executable

and readable.

```
res<-data.frame(res)
rownames(res)<-c("1":"9")
names(res)<-c("Age","Slope","Pr(slope)","R-squared","Adjusted_R-squared",
              "F-statistic_p_value")
print(res)
```

```
##   Age        Slope    Pr(slope) R-squared Adjusted_R-squared
## 1   1 0.0014659849 3.689370e-29 0.8871414          0.8851955
## 2   2 0.0010214916 2.291284e-29 0.8889762          0.8870620
## 3   3 0.0007135058 8.088674e-32 0.9085855          0.9070094
## 4   4 0.0005386710 8.657857e-37 0.9383615          0.9372987
## 5   5 0.0004346941 3.733806e-36 0.9351791          0.9340615
## 6   6 0.0003580306 5.946366e-33 0.9164423          0.9150017
## 7   7 0.0002681619 9.986492e-30 0.8921041          0.8902439
## 8   8 0.0001792081 3.859873e-27 0.8675609          0.8652775
## 9   9 0.0001148014 1.284204e-23 0.8249733          0.8219556
##   F-statistic_p_value
## 1        3.689370e-29
## 2        2.291284e-29
## 3        8.088674e-32
## 4        8.657857e-37
## 5        3.733806e-36
## 6        5.946366e-33
## 7        9.986492e-30
## 8        3.859873e-27
## 9        1.284204e-23
```

Finally, print the most affected age group.

```
me_agegroup<-res[which.max(res$Slope),]$Age
print(me_agegroup)
```

```
## [1] 1
```

After the linear regression, the curve with people aged from 25 to 29 has the highest slope, and we may consider that this age group is most affected by opioid use.

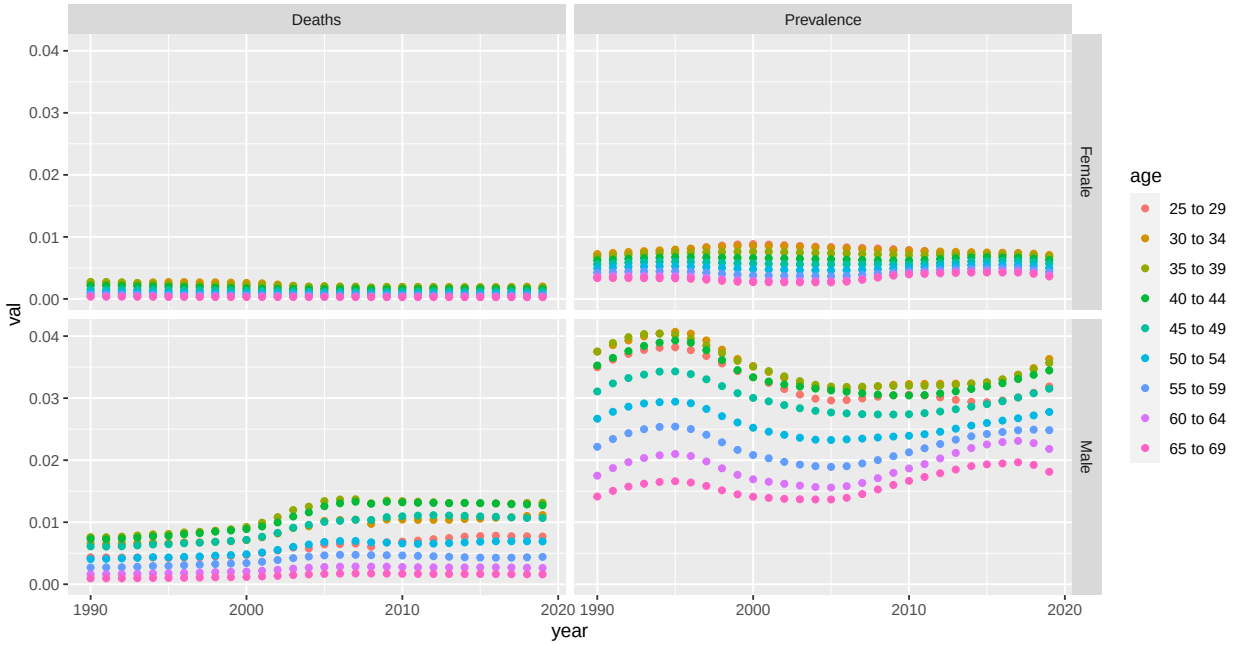
Part 2: Ask your own question

Our investigation to our own question:

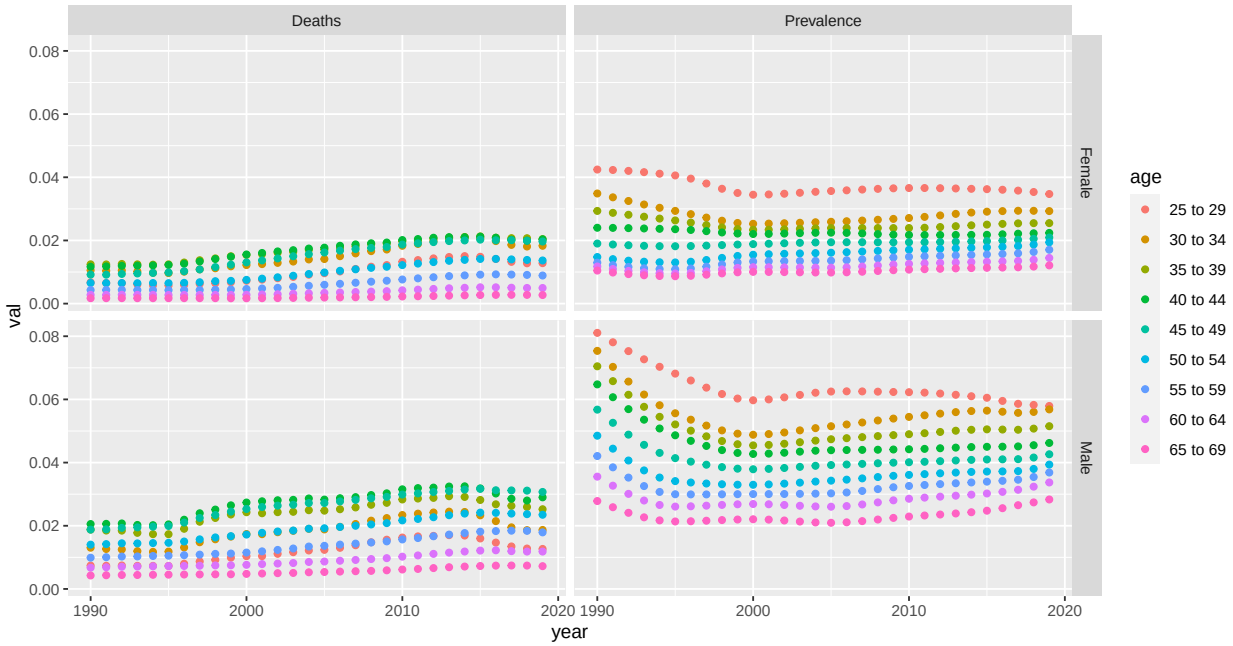
At first, we were curious about how has the prevalence and death rate of alcohol-related disorders changed over the years respectively in different regions around the world? And we decided to plot them for visualization.

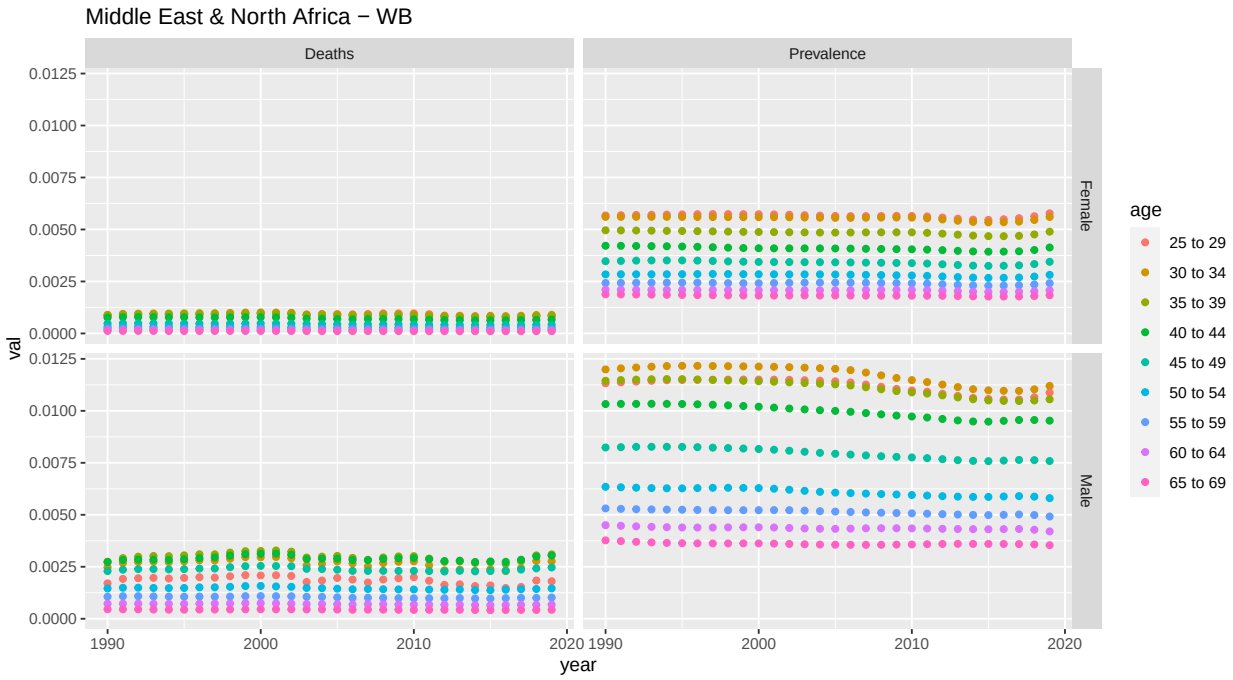
```
locations<-unique(dt$location)
data_al<-dt[which(dt$cause=="Alcohol use disorders"),]
for(loc in locations){
  sub_data<-data_al[which(data_al$location==loc),]
  p0<-ggplot(data=sub_data,aes(x=year,y=val))+
    geom_point(aes(color=age))+
    facet_grid(sex~measure)+
    ggtitle(loc)
  print(p0)
}
```

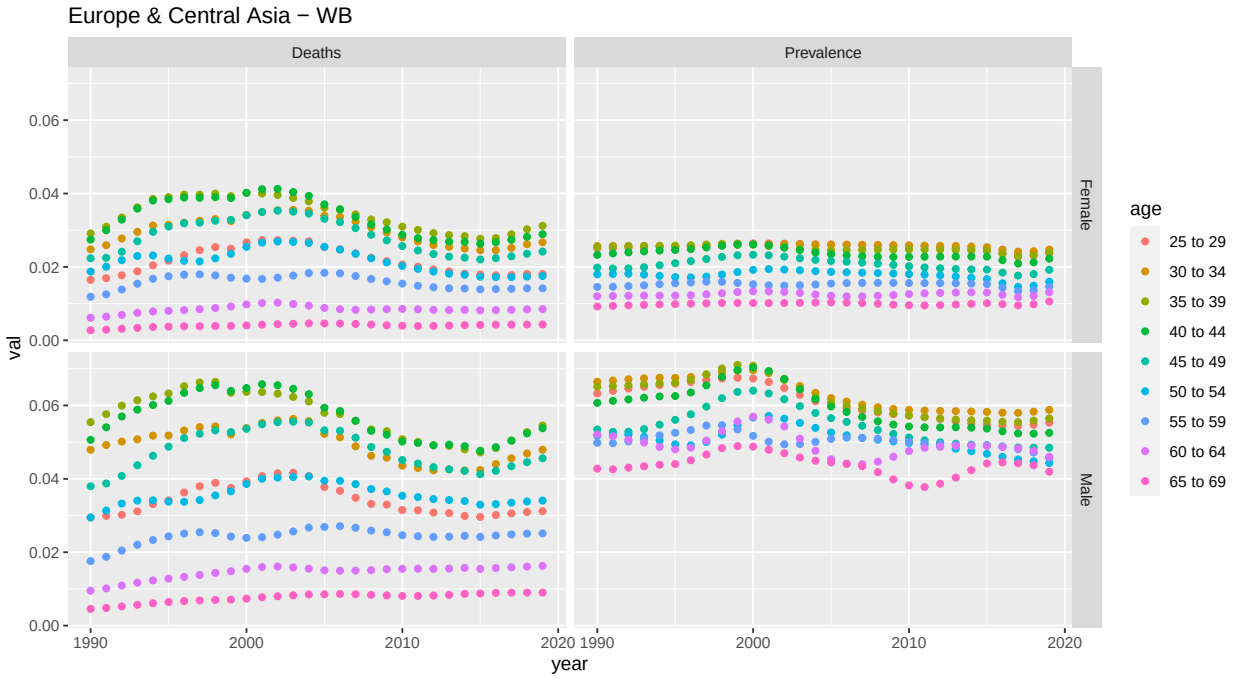
East Asia & Pacific − WB

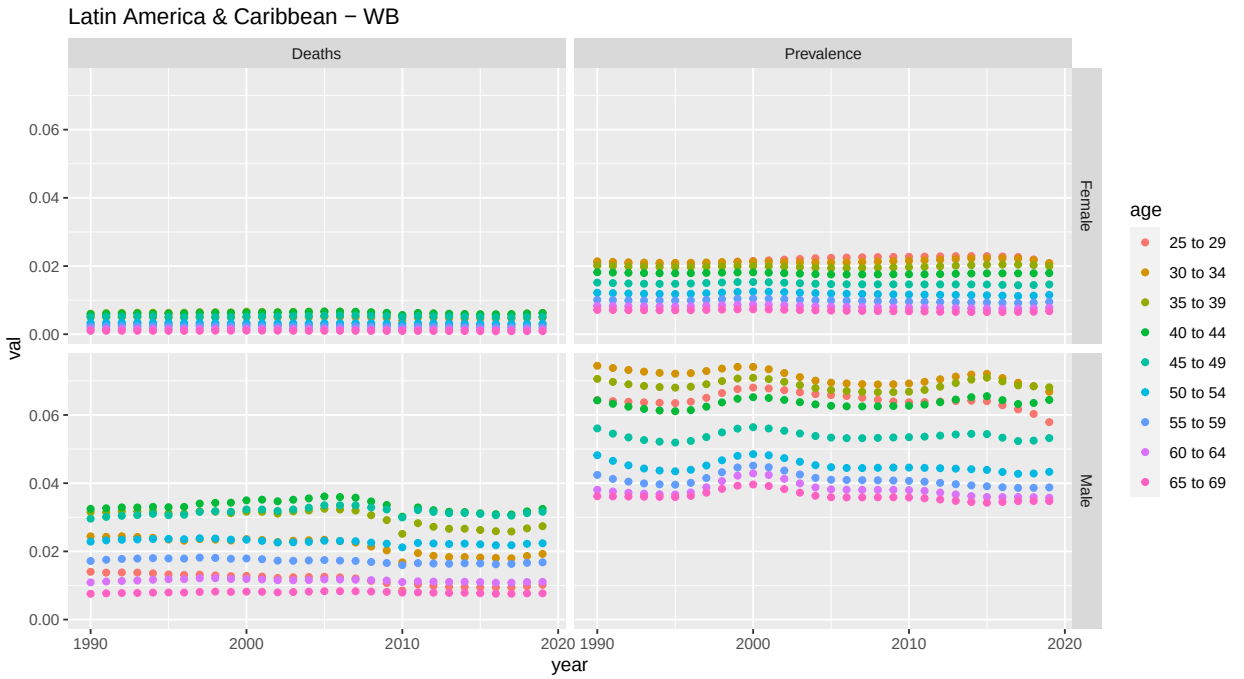


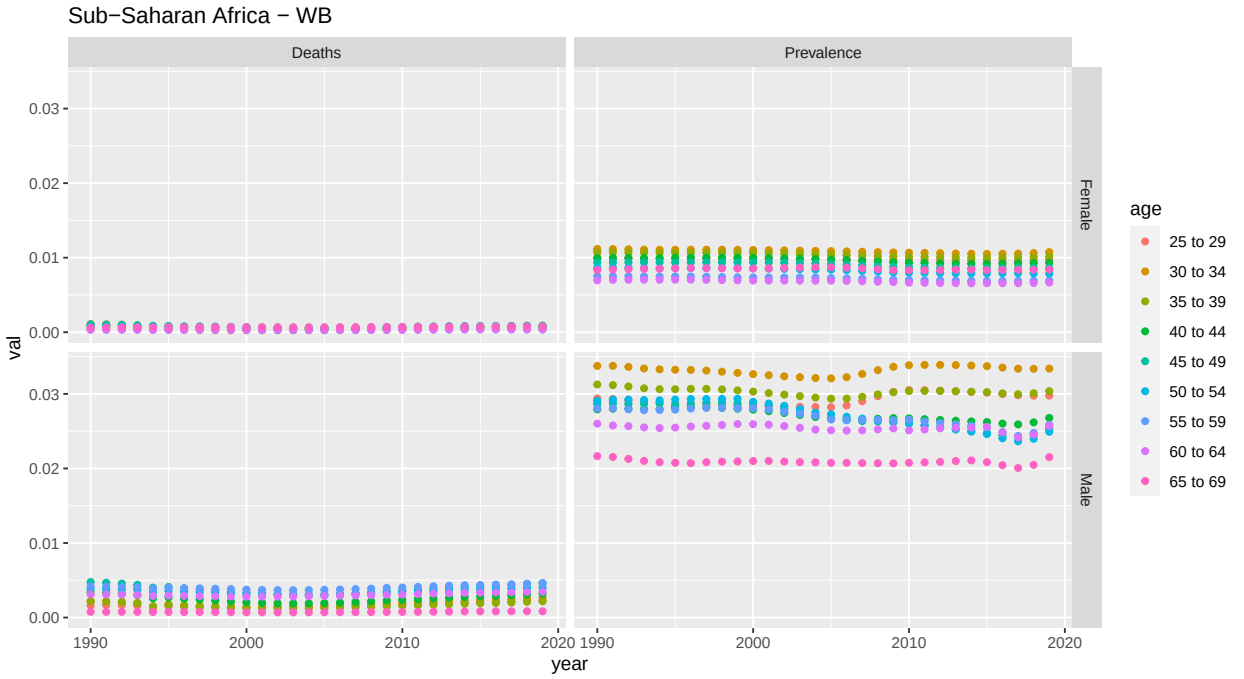
North America

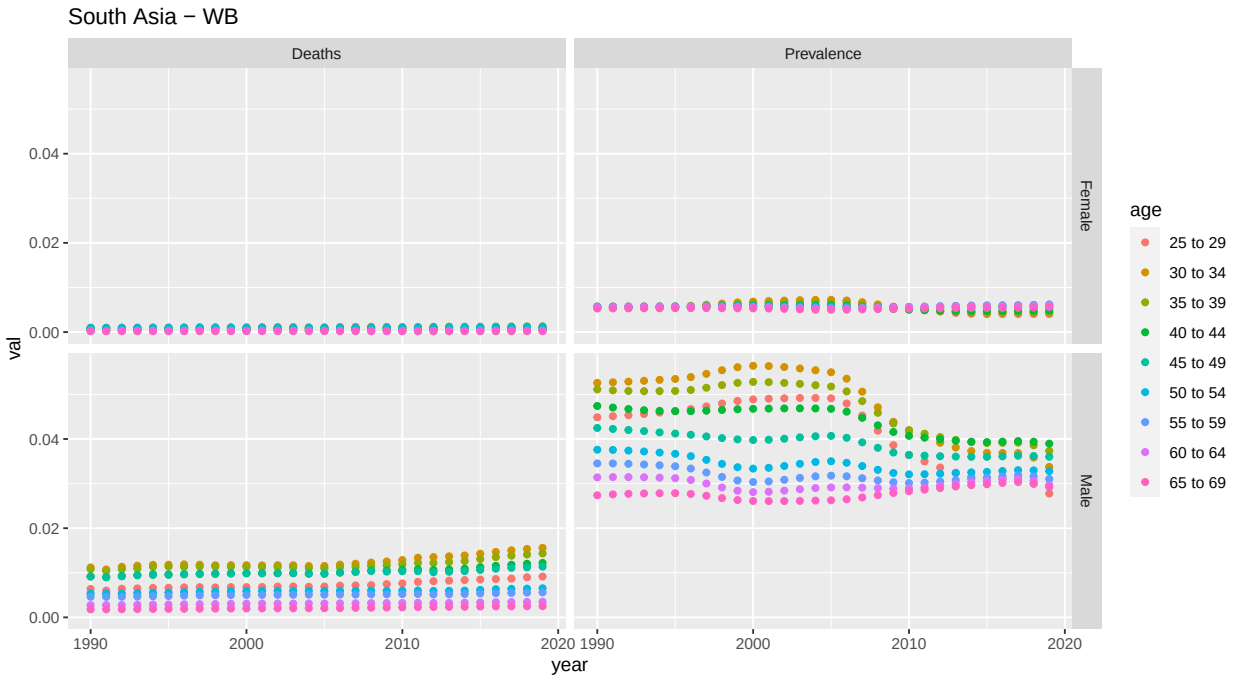



Middle East & North Africa − WB
Deaths
Prevalence
Female
Male
val
year
age
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69


Europe & Central Asia − WB
Deaths
Prevalence
Female
Male
val
year
age
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

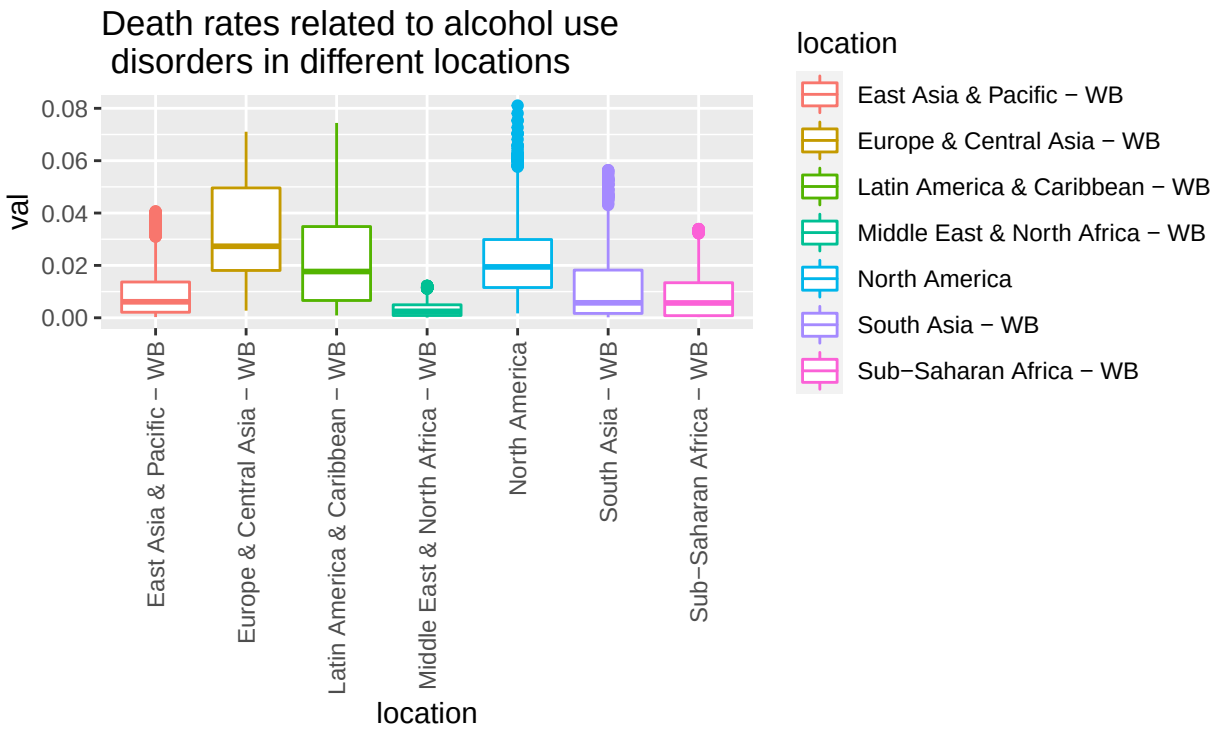

Latin America & Caribbean − WB
Deaths
Prevalence
Female
Male
val
year
age
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
65 to 69


Sub−Saharan Africa − WB
Deaths
Prevalence
Female
Male
val
year
age
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



First scanning from the plot, it gave us a feeling that the alcohol-related death rate in Europe and Central Asia is markedly higher than the other 6 regions. Thus, we plot out the data with the comparison between alcohol-related death rates in different locations.

```
data_al_d<-dt[which(dt$cause=="Alcohol use disorders",dt$measure=="Deaths"),]
ggplot(data_al_d)+geom_boxplot(aes(location,val,color=location))+
    theme(axis.text.x = element_text(angle=90, hjust=1, vjust=.5))+
    ggtitle("Death rates related to alcohol use \n disorders in different locations")
```



The new plot comparing death rates from different locations also seems to tell that the alcohol-related death rate in Europe and Central Asia is markedly higher than the other 6 regions.

Therefore, we decided to do a hypothesis test to see whether this is true.

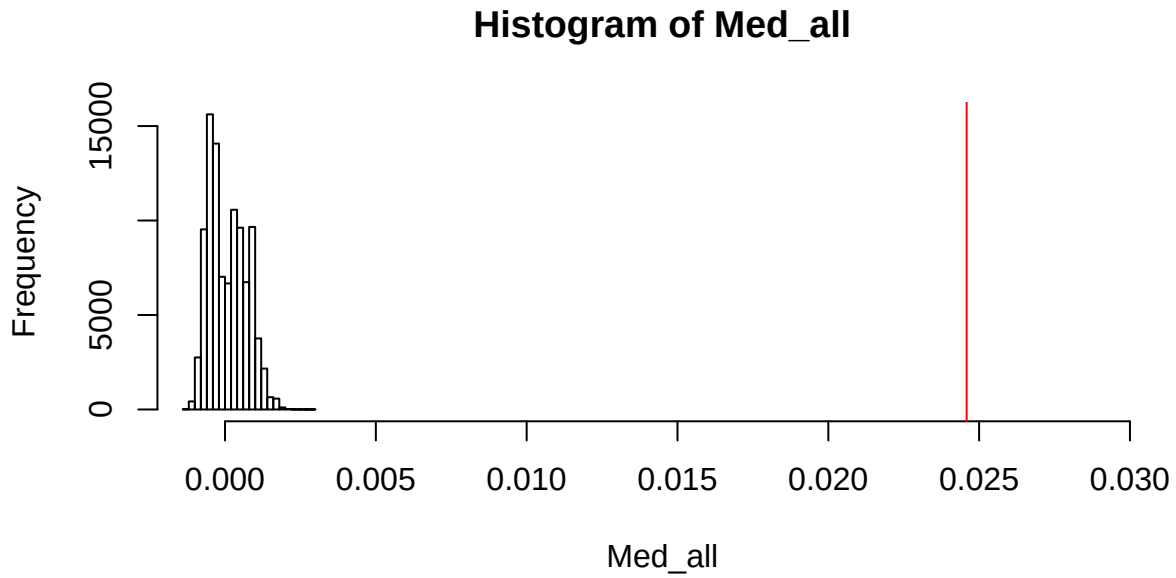
- Null hypothesis(H0): there is no difference between the death rate in Europe and Central Asia and the other 6 regions.
- Alternative hypothesis(H1): the death rate in Europe and Central Asia is higher than that in the other 6 regions of the world.

First, we try to use bootstrapping to get the p-value

```
no_ECA<-dt[which(dt$location!="Europe & Central Asia - WB"&
                   dt$measure=="Deaths"&dt$cause=="Alcohol use disorders"),]
ECA<-dt[which(dt$location=="Europe & Central Asia - WB"&
                 dt$measure=="Deaths"&dt$cause=="Alcohol use disorders"),]
pool<-rbind(ECA,no_ECA)
Med_all<-c()
med<-median(ECA$val)-median(no_ECA$val)
for(i in 1:100000){
  data_temp<-pool
  data_temp$val<-sample(pool$val,nrow(data_temp),replace = FALSE)
  eca<-data_temp[which(data_temp$location=="Europe & Central Asia - WB"),"val"]
  noEca<-data_temp[which(data_temp$location!="Europe & Central Asia - WB"),"val"]
  med_temp<-median(eca)-median(noEca)
  Med_all<-c(Med_all,med_temp)
}
p<-sum(Med_all>med)/100000
print(p)
```

```
## [1] 0
```

```
hist(Med_all,xlim = c(-0.001,0.03))
abline(v=med,col="red")
```



P in bootstrapping is 0, maybe the real p-value is too small to be simulated, therefore we want to use a t-test to see the p-value. And firstly we need to test the assumptions for t tets (i.e. Normal distribution, random sampling, independent mean, and errors).

```
shapiro.test(ECA$val)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ECA$val
## W = 0.96727, p-value = 1.313e-09
```

```
shapiro.test(no_ECA$val)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  no_ECA$val
## W = 0.74598, p-value < 2.2e-16
```

The p values for two samples are all smaller than 0.05, which means they are not normal-distributed, so we cannot use a t-test and we decided to use the Wilcoxon test (non-parametric test) instead.

```
var.test(ECA$val,no_ECA$val)
```

```
##
##  F test to compare two variances
##
## data:  ECA$val and no_ECA$val
## F = 4.3075, num df = 539, denom df = 3239, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  3.797383 4.915601
## sample estimates:
## ratio of variances
##           4.307461
```

```
wilcox.test(ECA$val,no_ECA$val,alternative = "greater",var.equal="F")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  ECA$val and no_ECA$val
## W = 1603620, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

The outcome of the Wilcoxon test shows that the p-value is smaller than 2.2e-16, and thus the null hypothesis is rejected, and the death rate in Europe and Central Asia is higher than that in other 6 regions in the world.

Our interpretation of the result:

Next we want to interpret our result from five different aspects:

Workload

- Concluding from history, usually, when a state is transferring to a higher stage of its social development, it faces the problems of an exceedingly high level of alcohol consumption, which may be associated with the heavy workload resulting from the rapid development. And there are many developing countries in central Asia, which all try hard to catch up with those developed countries. Moreover, although lots of

countries in Europe have already been developed countries, they also try to further develop themselves and compete with some more powerful countries. Therefore, they also face the challenge from both other developed countries and all developing countries.

Culture and lifestyle

- The trends of alcohol consumption are largely determined by western culture and lifestyle. In the countries heavily influenced by a ruling empire (so-called dependent territories, which are not uncommon in Central Asia), the consumption level is rather high, except for those regions which have special local customs and traditions that abandon drinking (e.g. those countries where Islam is a major religion).

Policy

- Government monopoly and business licensing system
  Government monopoly and business licensing system aim to provide a controllable system of alcohol business, thus better limiting the size of the alcohol industry. Based on the Global Status Report: Alcohol Policy from WHO, there are 12% of investigated countries that didn't apply any government monopoly and business licensing system, and most of which are European countries.
- Restrictions on the retail of alcohol
  To set restrictions on the retail of alcohol, there are mainly three parts, which are restrictions set on duration, restrictions set on the sites, and restrictions set on the density of the places for alcohol sales. Based on the Global Status Report: Alcohol Policy from WHO, the American Region has the highest amount of performed restrictions on duration and sites, followed by the Southeast Asia Region and West Pacific Region. On the contrary, the European and African regions have fewer restrictions. Lack of restriction together with a better economic condition in the European Region may account for the high alcohol consumption.

Weather

- Weather conditions might be another factor for alcohol consumption. Populations in the cold and moist region will be more likely to consume a larger amount of alcohol because alcohol can accelerate blood circulation and create an illusion of 'warming' the body. For example, Russian has an institution of consuming vodka to dispel chilliness.

Access of Alcohol

- The easy access to alcoholic beverages is one of the reasons for the high alcohol consumption and the subsequent high death rate in Europe and Central Asia. Alcohol prices, economic conditions, and taxation level of alcohol production result in regional differences in alcohol consumption. The alcohol prices of European and Central Asia are either low or at a medium level (Figure 1), which promotes the consumption of alcoholic beverages. In addition, because of the advanced economic conditions and low tax demands in the European region, people in Europe and Central Asia have easy access to alcohol.

Further Validation

- To further check the credibility of our result, we searched for the alcohol consumption data on the website of the World Health Organization (WHO). From their report, it can be concluded that the percentage of current drinkers in 2016 was highest in European Region (EUR) (59.9%) (Figure 2). The alcohol per capita consumption (APC) was relatively high in EUR (around 10 liters) from 2000 to 2016 (Figure 3).

Additional Part: A previous question we've though of and the reasons why it was abandoned

- When we focused on part2, we first came up with an idea to find the factor that affects the death/prevalence rate most, since the dataset provided us with several factors like age group, location, and so on. However, when we tried to perform the MLR (Multiple Linear Regression) after the verification of assumptions of the MLR, the result showed that all factors influenced the death/prevalence rate significantly, which meant we can not tell the most influential factor. Moreover, geographic location cannot be simply regarded as a single factor like gender or age group, because factors such as weather and economics will be different as location changes. For example, the difference between North America and Europe could not be the same as the difference between North America and Africa. In other words, the effect of location is not linear like age group, using linear regression might be meaningless.

References

World Health Organization. (2018) *Global status report on alcohol and health 2018*. Available at: https://apps.who.int/iris/bitstream/handle/10665/274603/9789241565639-eng.pdf?ua=1 [Accessed 05/05/2021].

Solov'ev. (2009) Worldwide Alcohol Production and Distribution. *Studies on Russian Economic Development*. Vol. 21, No. 4, pp. 411–425.