ADS2 Group8 ICA

2024-04-07

Part 1: Exploring the data

```
# First we import from the csv file
rawdata <- read.csv("substance_use.csv")

# Check if the raw data is clean
anyNA(rawdata)
```

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

```
anyDuplicated(rawdata)
```

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

The data has no NA and no duplication

```
# We can also check the size and column information of the rawdata
# The column information is clearly explained in the ICA guidance
dim(rawdata)
```

```
## [1] 15120    10
```

```
colnames(rawdata)
```

```
##  [1] "measure"  "location" "sex"      "age"      "cause"    "metric"
##  [7] "year"     "val"      "upper"    "lower"
```

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

```
# We select data with following information: year 2019, alcohol-related, deaths,
# gender male, age 40-44 and we want to research the difference between regions
subdf <- rawdata[which(rawdata$year == 2019
                       & rawdata$cause == "Alcohol use disorders"
                       & rawdata$measure == "Deaths"
                       & rawdata$sex == "Male"
                       & rawdata$age == "40 to 44"),]
```

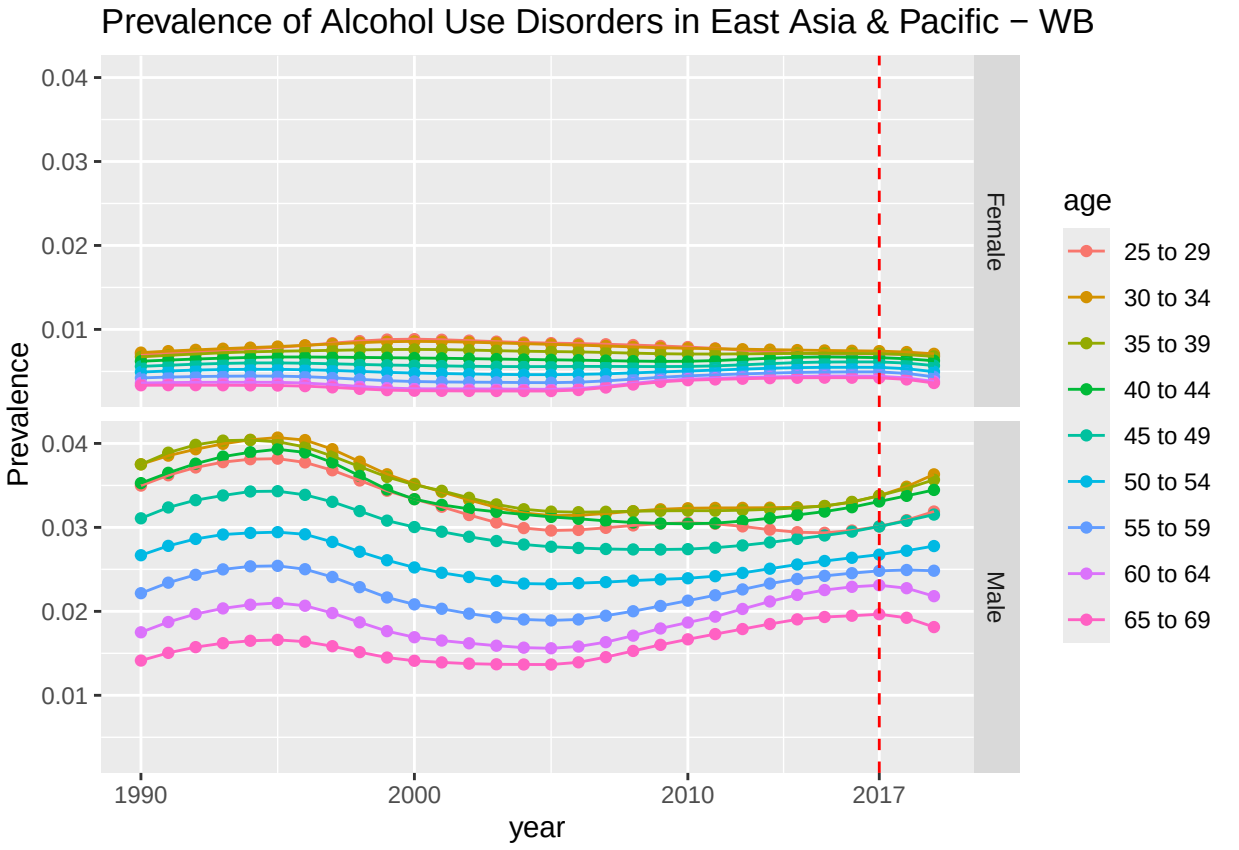
```r
# We sort the subset data by value to get the highest rate region
sorted <- subdf[order(subdf$val, decreasing = TRUE),]
highest_region <- sorted$location[1]
print(highest_region)
```

```
## [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? Is there a difference between men and women?**

```r
# We select data with following information: East Asia & Pacific - WB, Alcohol use
# disorders, Prevalence and we want to research the difference between age groups and sex
EastAsiaPacific_Alcohol <- rawdata[which(rawdata$location == "East Asia & Pacific - WB"
                                         & rawdata$cause == "Alcohol use disorders"
                                         & rawdata$measure == "Prevalence"),]
# To check the content in each column for selection we can use
# unique(rawdata$<colname>)

# We extract the values for each group at year = 2017 for sex = male
label_data <- EastAsiaPacific_Alcohol[which(EastAsiaPacific_Alcohol$year
                                            == 2017 & EastAsiaPacific_Alcohol$sex
                                            == 'male'),]

# We plot the data, grouping them by age and sex
ggplot(EastAsiaPacific_Alcohol, aes(x = year, y = val, color = age)) +
  geom_point() +
  geom_line() +
  ylab("Prevalence") +
  ggtitle("Prevalence of Alcohol Use Disorders in East Asia & Pacific - WB") +
  geom_vline(aes(xintercept = 2017), linetype = "dashed",   color = "red") +
  facet_grid(vars(sex)) +
  scale_x_continuous(breaks = sort(c(seq(min(EastAsiaPacific_Alcohol$year),
                                         max(EastAsiaPacific_Alcohol$year), 10),2017)))
```



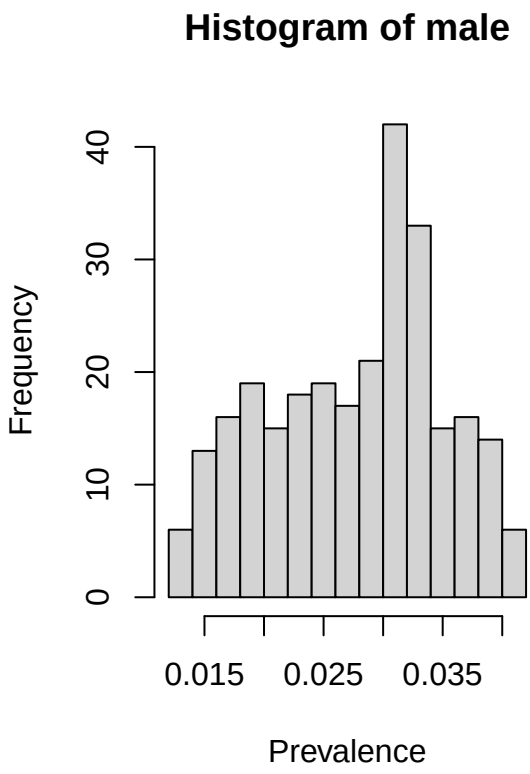
The plot shows us that the prevalence of alcohol use disorders.

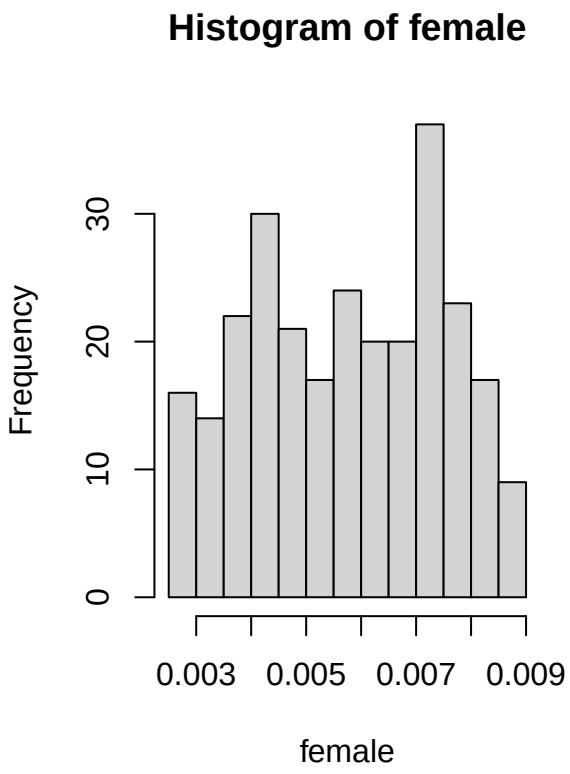
In women, the prevalence did not change much over time. Younger age groups tend to have higher prevalence compare to older age groups.
In men, younger age groups also tend to have higher prevalence compare to older age groups. All age groups experienced similar changes before 2017, which is rise, drop and then rise again. After 2017, age 25-59 still goes up while age 60-69 went down.
Overall, though prevalence in men went through ups and downs, they did not show significant trend of increasing or decreasing.

Given that the y-axis is uniform across both plots, there is a noticeable difference between men and women. It appears that the values for men are consistently higher than those for women. Specifically, the prevalence for men always exceeds 0.01, while for women, it remains below 0.01, but we better **do a statistical test to demonstrate the difference**.

H0: There is no difference between men and women.
Ha: There is significant difference between men and women.

```
# To test difference between men and women, we have to take out the information
male <- EastAsiaPacific_Alcohol[which(EastAsiaPacific_Alcohol$sex == "Male"),]$val
female <- EastAsiaPacific_Alcohol[which(EastAsiaPacific_Alcohol$sex == "Female"),]$val

# This is a two-sample test, check the samples can be view as normally distributed or not?
# Test whether it fits the normal distribution
par(mfrow = c(1,2))
hist(male, xlab = "Prevalence")
hist(female)
```






```
shapiro.test(male)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male
## W = 0.96139, p-value = 1.276e-06
```

```
shapiro.test(female)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female
## W = 0.96006, p-value = 8.645e-07
```

From the histograms, **apparently we cannot view the sample as normally distributed**. Therefore we did Wilcox test, here we have considered to use paired test because otherwise we are ignoring age and year differences. However, as the men and women are **not the same population** under different condition, we gave up paired test.

```
wilcox.test(x = male, y = female, alternative = "two.sided", paired = FALSE)
```

```
##
```

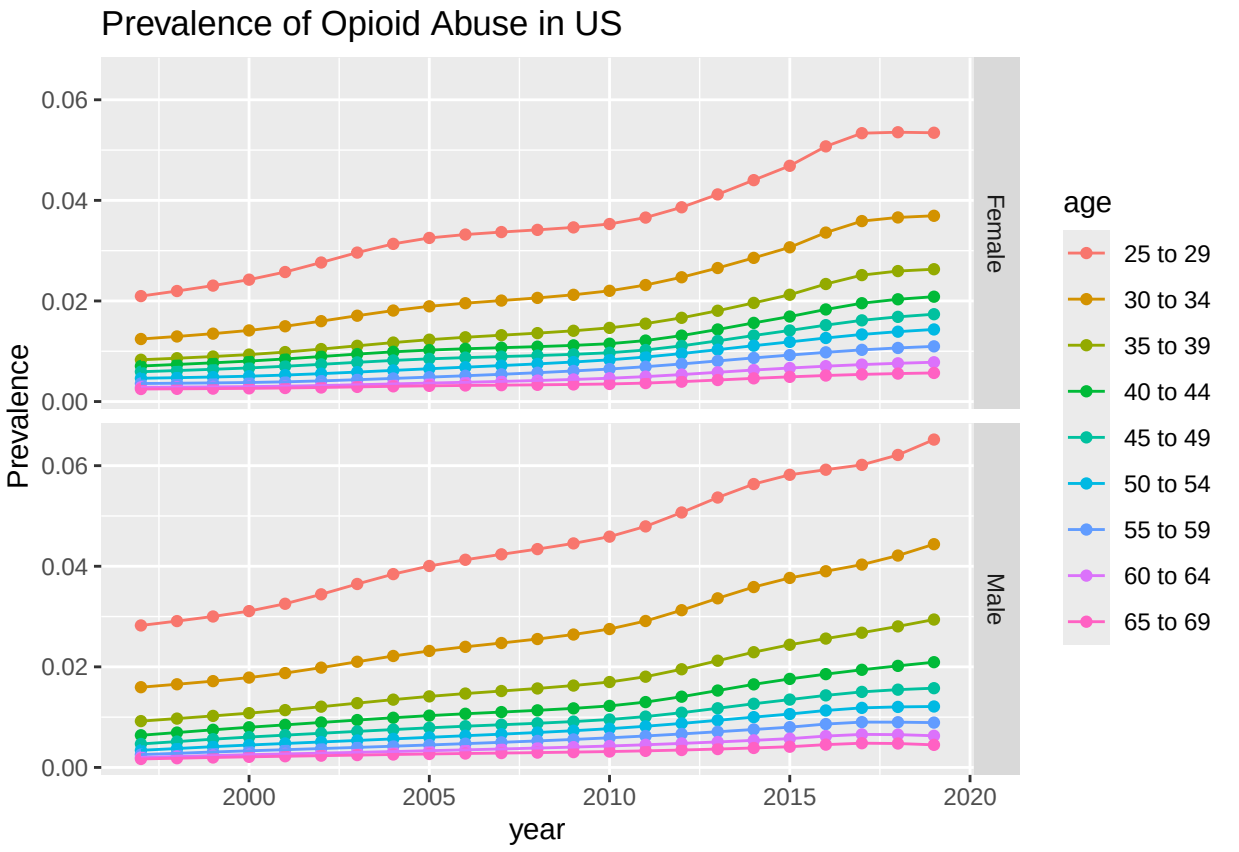
```
##  Wilcoxon rank sum test with continuity correction
##
## data:  male and female
## W = 72900, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Given that the p-value is less than 0.05, we reject the null hypothesis. This indicates a significant difference in the prevalence of alcohol-related diseases between men and women in the East Asia and Pacific region.

- **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?**

Given the ambiguity of the term **the late 1990s**, as outlined in the question, we choose to use **1997** as the starting point for our research, rather than selecting 1998 or 1999. This approach expands our time range, allowing us to **cover a more extensive scope and lend greater credibility to our findings.**

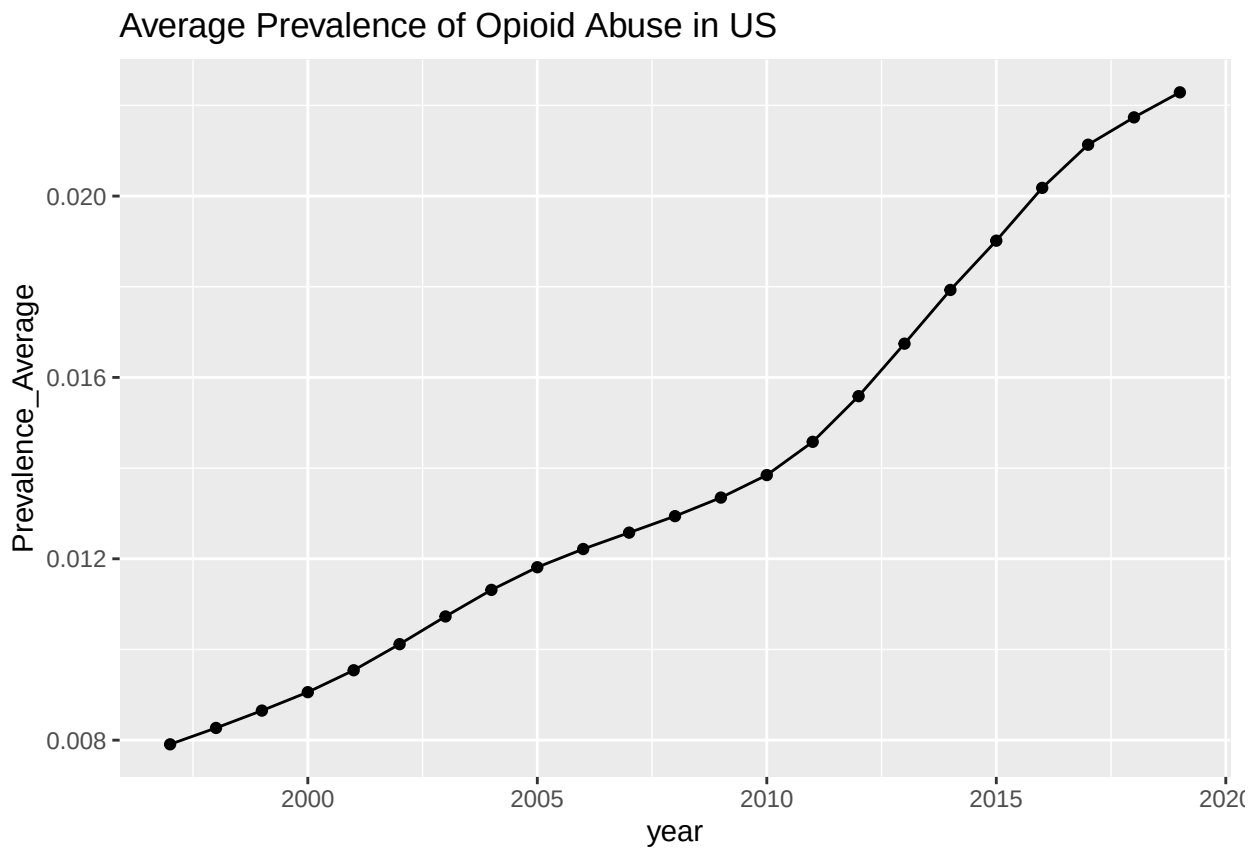
```r
# We select the data after 1997 again
opioidAbuse_In_US <- rawdata[which(rawdata$location == "North America"
                                   & rawdata$cause == "Opioid use disorders"
                                   & rawdata$measure == "Prevalence"
                                   & rawdata$year >= 1997),]
opioidAbuse_In_US <- opioidAbuse_In_US[order(opioidAbuse_In_US$year),]

# We first plot the data to check it visually
ggplot(opioidAbuse_In_US, aes(x = year, y = val, color = age))+
  geom_point() +
  geom_line() +
  ylab("Prevalence") +
  facet_grid(vars(sex)) +
  ggtitle("Prevalence of Opioid Abuse in US")+
  scale_x_continuous()
```



The plot illustrates a rising prevalence of opioid use disorders in the years spanning 1997-2019, with younger age groups consistently exhibiting a higher prevalence rate.

As there are too many groups if we consider age and sex, we **average all the prevalence**, regardless of age and sex, to represent the raw data. This method might cause some errors as we ignored so many dimensions of the information. Therefore we decided to draw a plot to see whether our processed data can represent the raw data.

```
prevalence_opioid_average <-  opioidAbuse_In_US %>% group_by(year) %>%
                              summarise(SummedValue = mean(val))
ggplot(prevalence_opioid_average, aes(x = year, y = SummedValue)) +
  geom_point() +
  geom_line() +
  ylab("Prevalence_Average")+
  ggtitle("Average Prevalence of Opioid Abuse in US")+
  scale_x_continuous()
```



After we average of all prevalence, we assume the new data can be used to represent the raw data.

As we only want to get the trend of the values, we use **Mann-Kendall trend test** (Mann-Kendall test is very similar to Pearson test,but instead of using raw data,it use rank as input)

H0: There is no trend of this data.
Ha: There is significant trend of this data.

```
mk.test(prevalence_opioid_average$SummedValue, continuity = TRUE)
```

```
##
##  Mann-Kendall trend test
##
## data:  prevalence_opioid_average$SummedValue
## z = 6.6554, n = 23, p-value = 2.825e-11
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##        S     varS      tau
##  253.000 1433.667    1.000
```

p-value is smaller than 0.05 means that we reject null hypothesis. Tau value is 1 telling us that the data is monotonically increasing. **We confirm an increase in the prevalence of opioid abuse disorders.**

We believe that the data from both sexes can be averaged to accurately represent both genders, **given the usual sex ratio of approximately 1:1, the result would not be effected by sample size of sexes**.

```r
prevalence_opioid_average_sex <-  opioidAbuse_In_US %>% group_by(year,age) %>%
                                  summarise(SummedValue = mean(val))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

In the plot, the majority of data points across all groups and sex appear to reveal a **monotonically increasing trend** over the course of the past three decades. Therefore, we elect to **measure the growth in prevalence by taking the difference between data from the first and final year**. Moreover, we've chosen to examine the absolute growth rate, as in this database, **relative growth rates can be significantly influenced when the base value is small**. Such magnitude-propelled bias might disguise the actual growth situation and thus, an examination of the absolute growth gives a more accurate reflection of the increase in prevalence.

```r
first_year <- head(sort(unique(opioidAbuse_In_US$year)), 1)
last_year <- tail(sort(unique(opioidAbuse_In_US$year)), 1)
# Then we calculate the difference and put them into a dataframe
diff_by_age <- data.frame(age = unique(opioidAbuse_In_US$age),
                          diff = NA)
for (i in diff_by_age$age) {
  first_year_data <- prevalence_opioid_average_sex$
    SummedValue[which(prevalence_opioid_average_sex$age == i &
                      prevalence_opioid_average_sex$year == first_year)]
  last_year_data <- prevalence_opioid_average_sex$
    SummedValue[which(prevalence_opioid_average_sex$age == i &
                      prevalence_opioid_average_sex$year == last_year)]
  diff_by_age$diff[which(diff_by_age$age == i)] <- last_year_data - first_year_data
}

# We sort the differences to get the most affected age group
diff_sorted <- diff_by_age[order(diff_by_age$diff, decreasing = TRUE),]
highest_age_group <- diff_sorted$age[1]
print(highest_age_group)
```

```
## [1] "25 to 29"
```

Based on the result, we believe the age group 25-29 is most affected.

Part 2: Ask your own question

Q:Whether there are sex differences in the prevalence of opioid abuse over time among 25-29 years old in the United States, and what models can be used to represent this relationship appropriately?

We mainly want to focus on the difference between gender, but as different locations and ages have different effects on the difference between men and women, we need to narrow the sample size.
We draw inspiration from the previous question and choose **'North American 25-29 year old'** as the sample to analyze the differences between men and women. We first observed the differences in the prevalence of drug abuse among people in different regions over time, and found that North America was most affected (The code is shown below), indicating that this part of the population is relatively significant for research, and a larger impact can better analyze the difference between genders statistically.

The most affected country

```
# Organize the data we need
prevalence_opioid_region_sum <- rawdata[which(rawdata$cause == "Opioid use disorders"
                                    & rawdata$measure=="Prevalence"),] %>%
  group_by(year,location) %>%
  summarise(SummedValue = sum(val))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
# Calculate to determine the most affected countries
first_year <- head(sort(unique(prevalence_opioid_region_sum$year)), 1)
last_year <- tail(sort(unique(prevalence_opioid_region_sum$year)), 1)

# Then we calculate the difference and put them into a dataframe
diff_by_location <- data.frame(location = unique(prevalence_opioid_region_sum$location),
                           diff = NA)

for (i in diff_by_location$location) {
  first_year_data <- prevalence_opioid_region_sum$
    SummedValue[which(prevalence_opioid_region_sum$location == i &
                      prevalence_opioid_region_sum$year == first_year)]
  last_year_data <- prevalence_opioid_region_sum$
    SummedValue[which(prevalence_opioid_region_sum$location == i &
                      prevalence_opioid_region_sum$year == last_year)]
  diff_by_location$diff[which(diff_by_location$location == i)] <- last_year_data -
                                                                  first_year_data
}

# We sort the differences to get the most affected location group
diff_sorted <- diff_by_location[order(diff_by_location$diff, decreasing = TRUE),]
highest_location_group <- diff_sorted$location[1]

print(highest_location_group)
```

```
## [1] "North America"
```

The most affected country is the United States.

Then we focused on the difference between gender. We first conduct linear regression analysis and evaluation separately for the two genders.

Data Cleaning

```
# Clean data: select North America, prevalence data, opioid use disorders as cause
# and remove missing values
data_clean <- rawdata %>%
  filter(location == "North America", measure == "Prevalence",
         cause == "Opioid use disorders", age == "25 to 29") %>%
  na.omit()
```

Correlation Analysis

```
# Select data for males and females
data_males <- subset(data_clean, sex == "Male")
data_females <- subset(data_clean, sex == "Female")

# Calculate the correlation of opioid use disorders prevalence over time in both genders
correlation_males <- cor.test(data_males$val, data_males$year, use = "complete.obs")
correlation_females <- cor.test(data_females$val, data_females$year, use = "complete.obs")

# Print the results of the correlation tests
print(correlation_males)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_males$val and data_males$year
## t = 52.545, df = 28, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9893298 0.9976301
## sample estimates:
##       cor
## 0.9949676
```

```
print(correlation_females)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data_females$val and data_females$year
## t = 28.855, df = 28, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9654430 0.9922524
## sample estimates:
##       cor
## 0.9835978
```

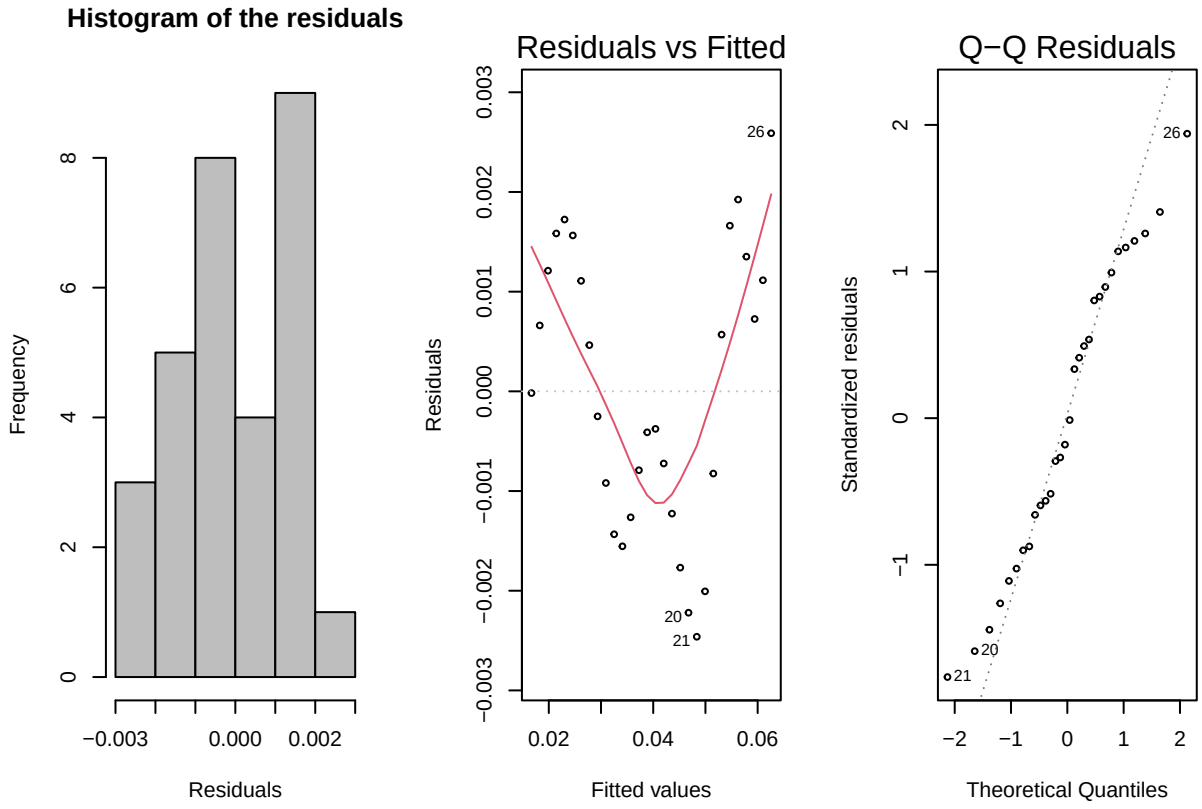
We got that the correlation of opioid use disorders prevalence and time for males is 0.9949676 while for female is 0.9835978. **Both values are very close to 1 and it can be said that both men and women conform to linear relationship between opioid use disorders prevalence and time**.

Model Building and Evaluation

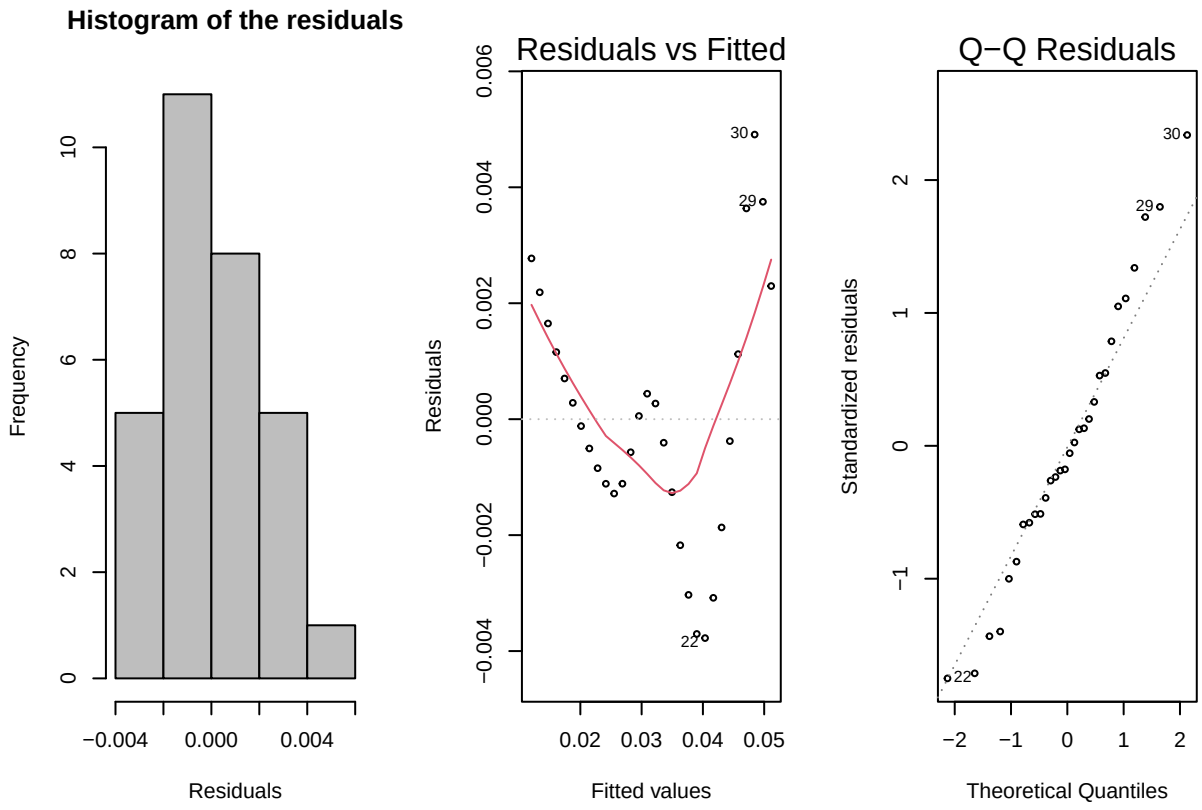
To verify the appropriateness of the linear fit, we need to check whether the residuals of the linear model follow a normal distribution.

```
# Build models
model_male <- lm(data_males$val ~ data_males$year)
```

```r
model_female <- lm(data_females$val ~ data_females$year)

# Evaluate models

#For male
par(mfrow = c(1, 3))
hist(residuals(model_male), breaks = 5, col = "gray",
     main = "Histogram of the residuals", xlab = "Residuals", cex = 0.6)
plot(model_male, which = c(1, 2), cex = 0.6)
```




```r
#For female
par(mfrow = c(1, 3))
hist(residuals(model_female), breaks = 5, col = "gray",
     main = "Histogram of the residuals", xlab = "Residuals", cex = 0.6)
plot(model_female, which = c(1, 2), cex = 0.6)
```

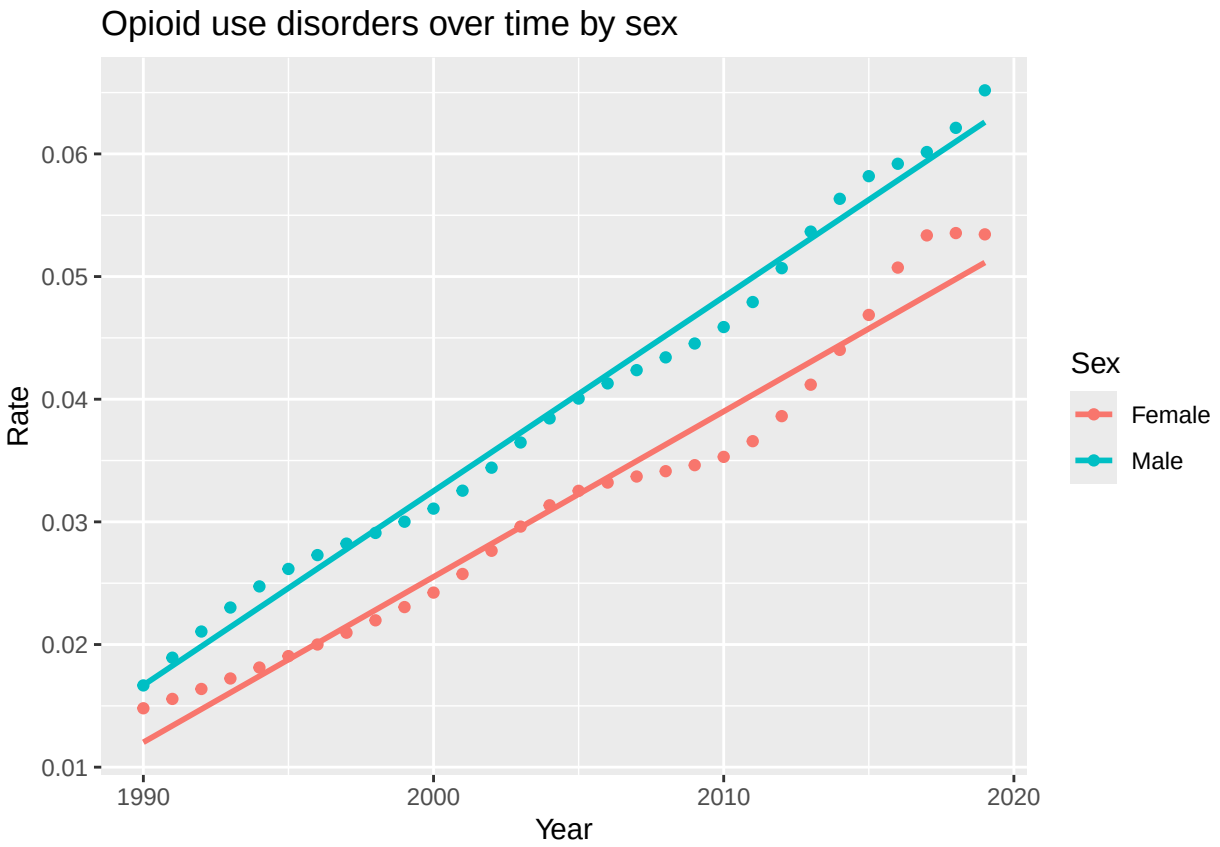


Here the sample size is small so the histogram is not very clear, but the QQ plot shows the normality well

Visualization

```
# Create scatter plot and regression line
ggplot(data_clean, aes(x=year, y=val, color=sex)) +
  geom_point() +
  geom_smooth(method=lm, se = FALSE) +
  labs(title="Opioid use disorders over time by sex", x="Year", y="Rate", color="Sex")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Notice that the prevalence rate for males is generally higher than that for females.

Calculate Z

Next, z value is calculated to determine whether the slopes of the two fitted lines are significantly different
H0: There is no difference between the two correlation coefficients.
Ha: There is significant difference between the two correlation coefficients.

```
# Extract regression coefficients and standard errors
beta1 <- summary(model_male)$coefficients["data_males$year", "Estimate"]
se_beta1 <- summary(model_male)$coefficients["data_males$year", "Std. Error"]
beta2 <- summary(model_female)$coefficients["data_females$year", "Estimate"]
se_beta2 <- summary(model_female)$coefficients["data_females$year", "Std. Error"]
# Calculate z-score
z <- (beta1 - beta2) / sqrt(se_beta1^2 + se_beta2^2)
print(z)
```

```
## [1] 4.21291
```

Here, **z is greater than 1.96 (significance level 0.05), indicating a significant difference between the two correlation coefficients, so we reject H0.** This suggests that there is a gender difference in the impact of drug addiction, with males being more affected and having a higher growth rate.

We try to build a model to illustrate this difference, starting with a multivariate regression analysis to build the model (below).

Multivariate Linear Regression Model

Since there are three variables here, which are gender, year, and prevalence rate, in order to explore the relationship among these three, we try to create a multivariate linear regression model to predict the prevalence rate. Note here that the two independent variables, age and sex, obviously have no collinearity relationship

```
# Convert the sex variable into a factor
data_clean$sex <- as.factor(data_clean$sex)
# Create a multivariate linear regression model
model_multi <- lm(val ~ year + sex, data = data_clean)
# Output the summary of the model
summary(model_multi)
```

```
##
## Call:
## lm(formula = val ~ year + sex, data = data_clean)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0045375 -0.0013916 -0.0002658  0.0014373  0.0044718
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.907e+00  6.340e-02  -45.85   <2e-16 ***
## year         1.466e-03  3.163e-05   46.35   <2e-16 ***
## sexMale      8.052e-03  5.475e-04   14.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002121 on 57 degrees of freedom
## Multiple R-squared:  0.9765, Adjusted R-squared:  0.9756
## F-statistic:  1182 on 2 and 57 DF,  p-value: < 2.2e-16
```

Using multivariate linear regression, we got the model:
y=-2.907+0.008052*sex+0.001466*x (Where sex(Male) = 1 and sex(Female) = 0)

In this issue, sex only has two options: male and female, which are not numerical data. To use it as an independent variable in multivariate analysis, it needs to be transformed into a dummy variable, encoded as 1 and 0. According to the results obtained from the previous z-value calculation, the hypothesis of no difference in the fitting growth rate between the sexes needs to be rejected. However, in this model, because it is a multivariate linear analysis, using sex as an independent variable can only change the intercept. Although the data obtained in this fitting is quite significant, a better model may be needed to allow the influence of sex to affect the slope of the fitted line, allowing the interaction between sex and year to impact the prediction of the disease rate, thereby conforming to the conclusions drawn from the previous z-value calculations. **This model should conform to y=a+b1*x+b2*sex+b3*x*sex.**

Better model

```
data_clean$sex <- as.factor(data_clean$sex)
model1 <- lm(val ~ year*sex, data = data_clean)
summary(model1)
```

```
##
## Call:
## lm(formula = val ~ year * sex, data = data_clean)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0037760 -0.0012348 -0.0001852  0.0011702  0.0049094
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.672e+00  7.883e-02 -33.899  < 2e-16 ***
## year          1.349e-03  3.932e-05  34.300  < 2e-16 ***
## sexMale      -4.616e-01  1.115e-01  -4.141 0.000118 ***
## year:sexMale  2.343e-04  5.561e-05   4.213 9.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001864 on 56 degrees of freedom
## Multiple R-squared:  0.9821, Adjusted R-squared:  0.9812
## F-statistic:  1026 on 3 and 56 DF,  p-value: < 2.2e-16
```

According to the results, this model indeed better fits the relationship among sex, time, and prevalence rate. (R-squared = 0.9821 > 0.9765)

So the better model is:
y=-2.672+0.001349*year-0.4616*sex+0.0002343*year*sex
(Where sex(Male) = 1 and sex(Female) = 0)