# ADS2 Group Exercise ICA

Group 6

2024-04-02

```
substance_use = read.csv("substance_use.csv")
```

## Part 1: Exploring the data

In this part, we will answer the questions given in the guidance.

### Question 1

#### Description

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

#### Method

First, we need to extract the data of the alcohol-related death rate among men aged 40-44.

We use pipeline operations to construct the code, which avoids excess intermediate variables, enhances readability and makes the logical relationship clearer.

```
highest_alcohol_deaths = substance_use %>%
  filter(measure == "Deaths", year == 2019, age == "40 to 44",
         sex == "Male", cause == "Alcohol use disorders") %>%
  # Filter the data
  select(location, val) %>%
  arrange(desc(val)) %>%
  # Sort by death rate in descending order
  top_n(1, val)

highest_alcohol_deaths
```

```
##                     location        val
## 1 Europe & Central Asia - WB 0.05379854
```

#### Conclusion

In 2019, Europe & Central Asia has the highest rate of alcohol-related deaths among men aged 40-44.

## Question 2

### Description

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?

### Method

First, we extract the data from different years, age groups and sex groups, and calculate the average prevalence rate in case of more than one values with identical attributes.

```
eap_alcohol_data = substance_use %>%
  filter(measure == "Prevalence", cause == "Alcohol use disorders",
         location == "East Asia & Pacific - WB") %>%
  select(year, age, sex, val)
```

Next, we visualize the data from different years, age groups and sex groups. We visualize the data in two ways, faceting the plots by age groups and by sex groups. We do not mix the data together by using the average number, because populations from different age groups and sex groups are different.

```
#eap_alcohol_trends1 = eap_alcohol_trends %>% paste0(.$age, " years old")
eap_alcohol_data1 = eap_alcohol_data
eap_alcohol_data1$age = paste0(eap_alcohol_data1$age, " years old")
ggplot(eap_alcohol_data1,
       aes(x = year, y = val, color = sex,
           group = interaction(sex, age))) +
  geom_line() +
  facet_wrap(~age, scales = 'free_y') +
  # Use the panel diagram to show the different age groups
  labs(x = "Year",
       y = "Prevalence (%)",
       color = "Sex") +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5)
        )
```

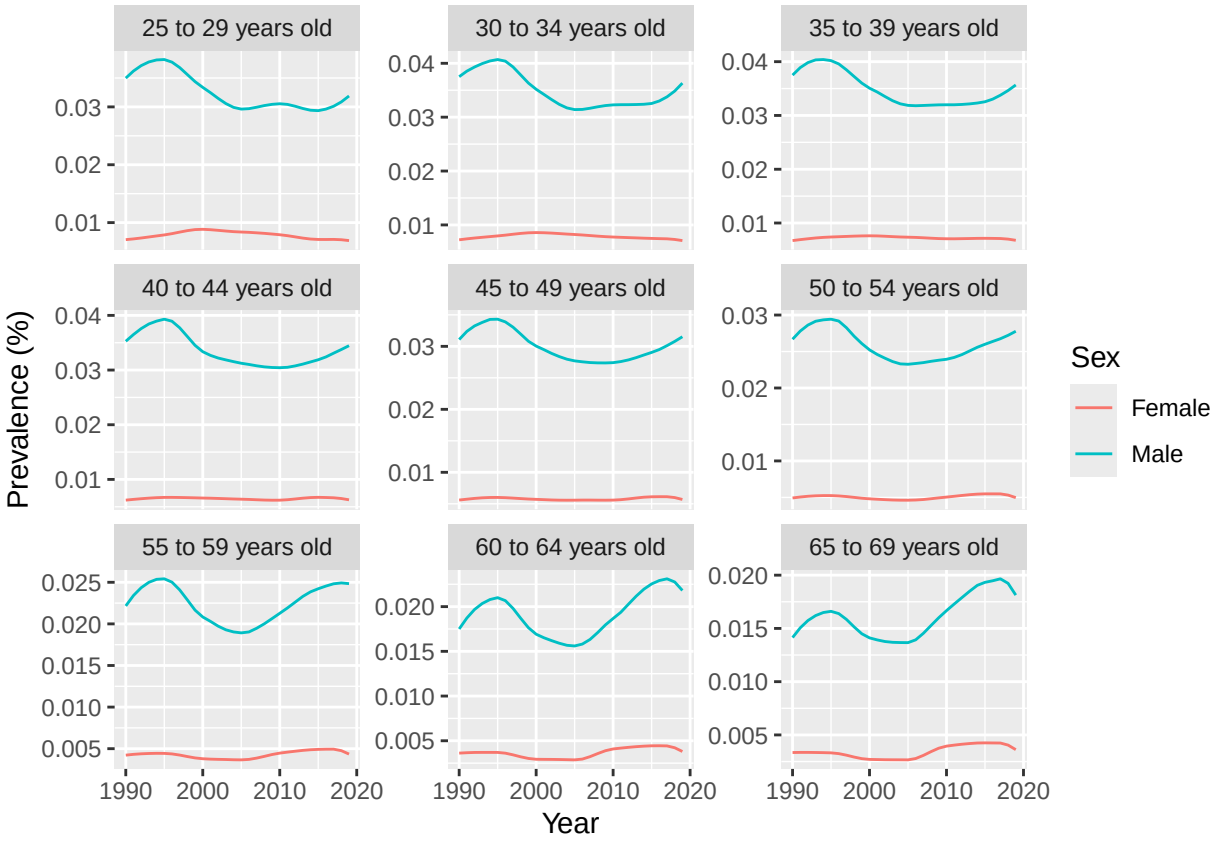



Figure 2.1: The trends of the alcohol-related disease prevalence in east Asia and Pacific. The plots in this figure are faceted by age groups, better for visually comparing the prevalence between the male and the female.

```
ggplot(eap_alcohol_data,
       aes(x = year, y = val, color = age,
           group = interaction(sex, age))) +
  geom_line() +
  geom_point(size = 0.5) +
  facet_wrap(~sex, scales = 'free_y') +
  # Use the panel diagram to show the different sex groups
  labs(x = "Year",
       y = "Prevalence (%)",
       color = "Age (years old)") +
  theme(#legend.position = "bottom",
        #legend.title.position = "top",
        legend.title = element_text(hjust = 0.5),
        plot.title = element_text(hjust = 0.5)
        )
```

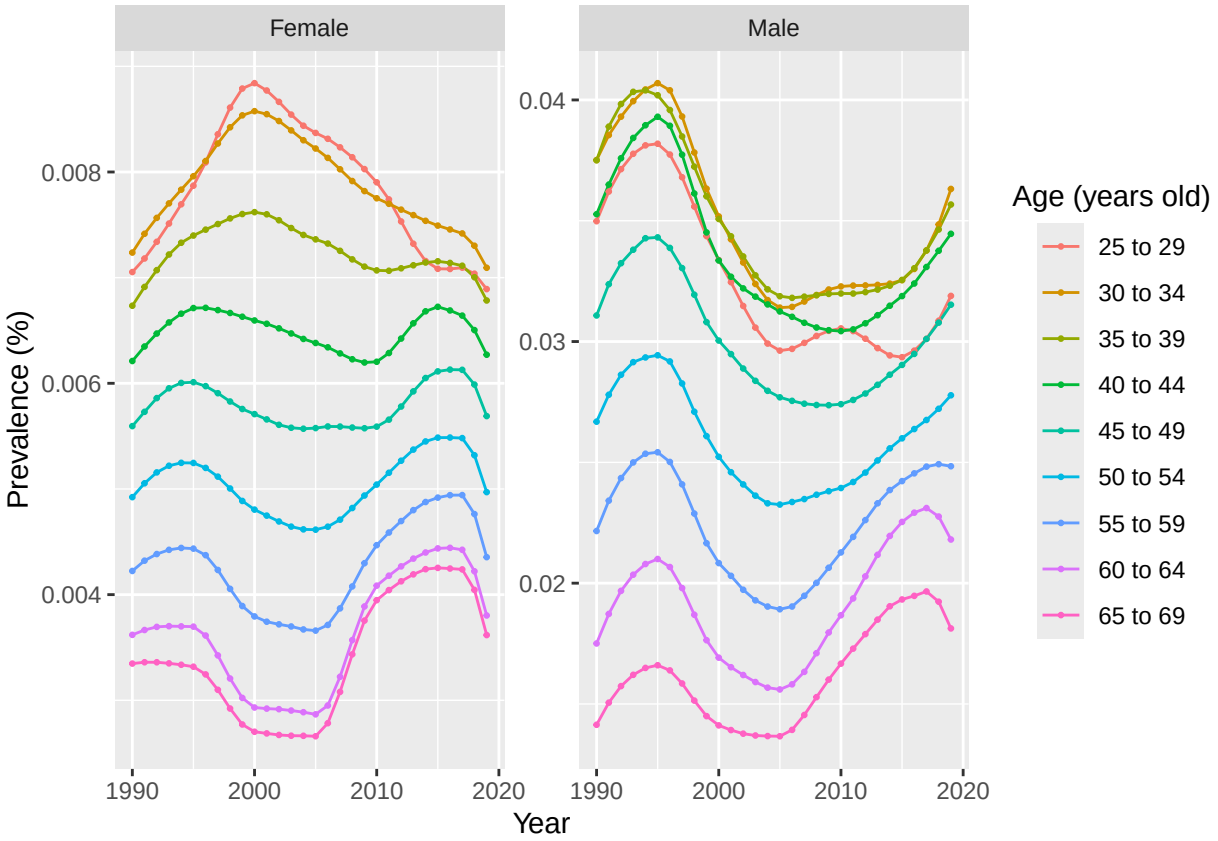



Figure 2.2: The trends of the alcohol-related disease prevalence in east Asia and Pacific. The plots in this figure are faceted by sex groups, better for visually comparing the prevalence between different age groups.

From the two plots, you can see how the prevalence of alcohol-related disease in the East Asia and Pacific region and in different age groups. Also, you can clearly identify different patterns of the prevalence between male and female.

In the following, we will validate the difference between male and female in a more statistical way.

First, we formulate the hypotheses.

- The null hypothesis (H0): There is no difference in the prevalence of different age groups between male and female.
- The alternative hypothesis (HA): There is difference in the prevalence of different age groups between male and female.

The data for male and female in the same year are correlated and paired (i.e. observational data under the same conditions), and we want to test the differences in prevalence between male and female in each year, so we will use the paired statistical test in the following. We use the Shapiro test to test the normality of the differences in prevalence between male and female.

- If the differences are normally distributed, we will perform the parametric Student's t-test to test the hypotheses.
- If the differences are not normally distributed, we will perform the non-parametric Wilcoxon test to test the hypotheses.

```r
eap_alcohol_data2 = eap_alcohol_data %>%
  group_by(age) %>%
  group_split() %>%
  map(~spread(., key = "sex", value = "val")) %>%
  # Convert to the wide format
  map(~{
    shapiro_p = shapiro.test(.$Male - .$Female)$p.value
    if(shapiro_p < 0.05){
      test_type = "Wilcoxon test"
      p_value = wilcox.test(.$Male, .$Female, paired = T)$p.value
    }
    else{
      test_type = "Student's t-test"
      p_value = t.test(.$Male, .$Female, paired = T)$p.value
    }
    tibble(shapiro_p = shapiro_p,
           test_type = test_type,
           p_value = p_value)
  }) %>%
  # perform statistical test in all age groups
  bind_rows() %>%
  mutate(age = unique(eap_alcohol_data$age), .before = shapiro_p)
eap_alcohol_data2
```

```
## # A tibble: 9 x 4
##   age      shapiro_p test_type           p_value
##   <chr>        <dbl> <chr>                 <dbl>
## 1 25 to 29  0.000199 Wilcoxon test 0.00000000186
## 2 30 to 34  0.00128  Wilcoxon test 0.00000000186
## 3 35 to 39  0.000252 Wilcoxon test 0.00000000186
## 4 40 to 44  0.000635 Wilcoxon test 0.00000000186
## 5 45 to 49  0.00239  Wilcoxon test 0.00000000186
## 6 50 to 54  0.00440  Wilcoxon test 0.00000000186
## 7 55 to 59  0.0289   Wilcoxon test 0.00000000186
## 8 60 to 64  0.0411   Wilcoxon test 0.00000000186
## 9 65 to 69  0.0185   Wilcoxon test 0.00000000186
```

### Conclusion

Of all the age groups, the p values from the statistical test are all less than $10^{-8}$. Therefore, we reject the null hypothesis(H0). There is sufficient evidence to support the conclusion that there are significant differences in the prevalence between the male and the female within each age group.

## Question 3

### Description

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?

### Method

There is no data named "United States" in the locations, because the locations are divided according to the world bank. Therefore, we extract the data from "North America" to represent the data from the United States.

According to the question, since we have the data from 1990 to 2019, we will separate the data into two parts "before the start of increasing prescription the opioid" (1990-1997) and "after the start of increasing prescription the opioid" (1998-2019). We will abbreviate the two periods to the "before" period and the "after" period in the following.

```
opioid_use_na = substance_use %>%
  filter(measure == "Prevalence",
         location == "North America",
         cause == "Opioid use disorders") %>%
  select(age, year, sex, val)
```

Next, we visualize the data from different years, age groups and sex groups. We visualize the data in two ways, faceting the plots by age groups and by sex groups.

```
opioid_use_na1 = opioid_use_na
opioid_use_na1$age = paste0(opioid_use_na1$age, " years old")
ggplot(opioid_use_na1,
       aes(x = year, y = val, color = sex,
           group = interaction(sex, age))) +
  geom_line() +
  geom_vline(xintercept = 1998, color = "grey") +
  facet_wrap(~age, scales = 'free_y') +
  # Use the panel diagram to show the different age groups
  labs(x = "Year",
       y = "Prevalence (%)",
       color = "Sex") +
  theme(legend.position = "right",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45)
        )
```

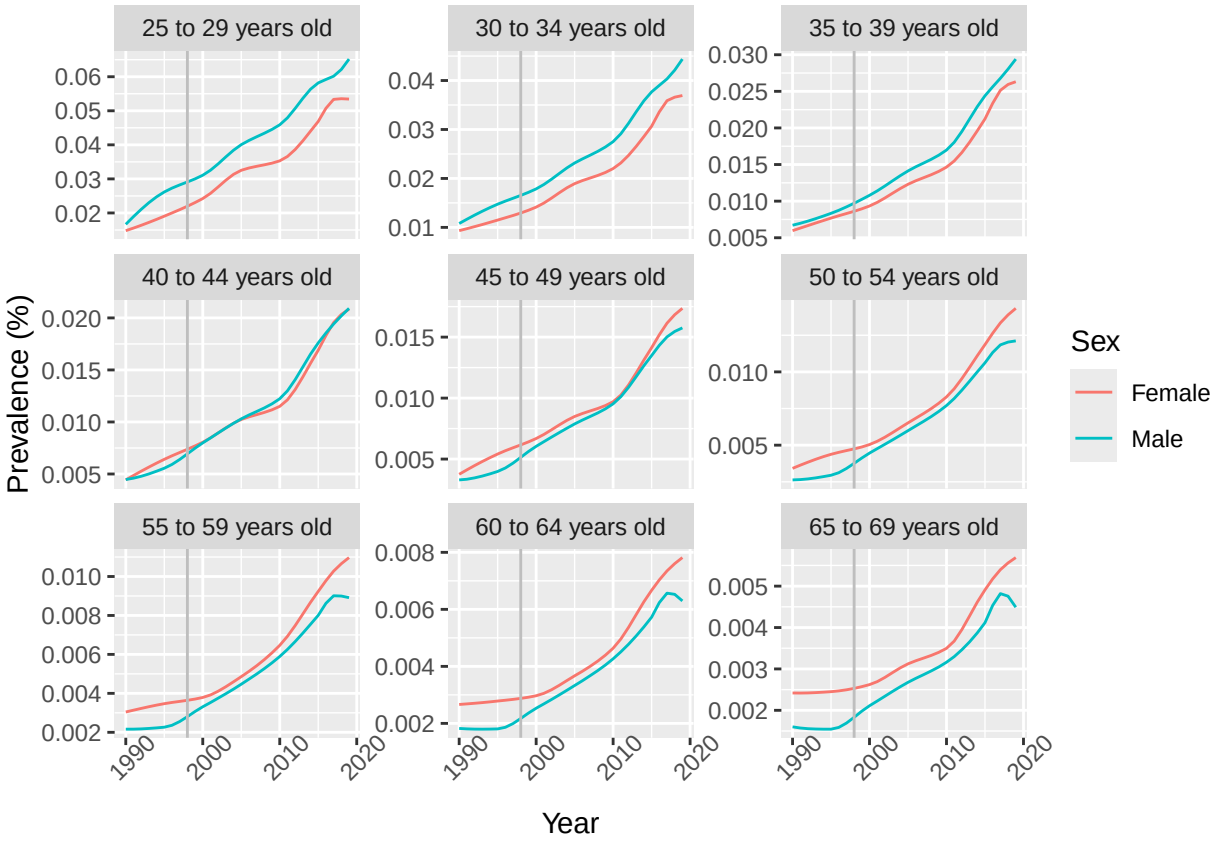



Figure 3.1: The trends of the opioid use related disease prevalence in North America. The plots in this figure are faceted by age groups, better for visually comparing the prevalence between the male and the female.

```
ggplot(opioid_use_na,
       aes(x = year, y = val, color = age,
           group = interaction(sex, age))) +
  geom_line() +
  geom_point(size = 0.5) +
  geom_vline(xintercept = 1998, color = "grey") +
  facet_wrap(~sex, scales = 'free_y') +
  # Use the panel diagram to show the different sex groups
  labs(x = "Year",
       y = "Average Prevalence (%)",
       color = "Age (years old)") +
  theme(#legend.position = "bottom",
        #legend.title.position = "top",
        legend.title = element_text(hjust = 0.5),
        plot.title = element_text(hjust = 0.5)
  )
```

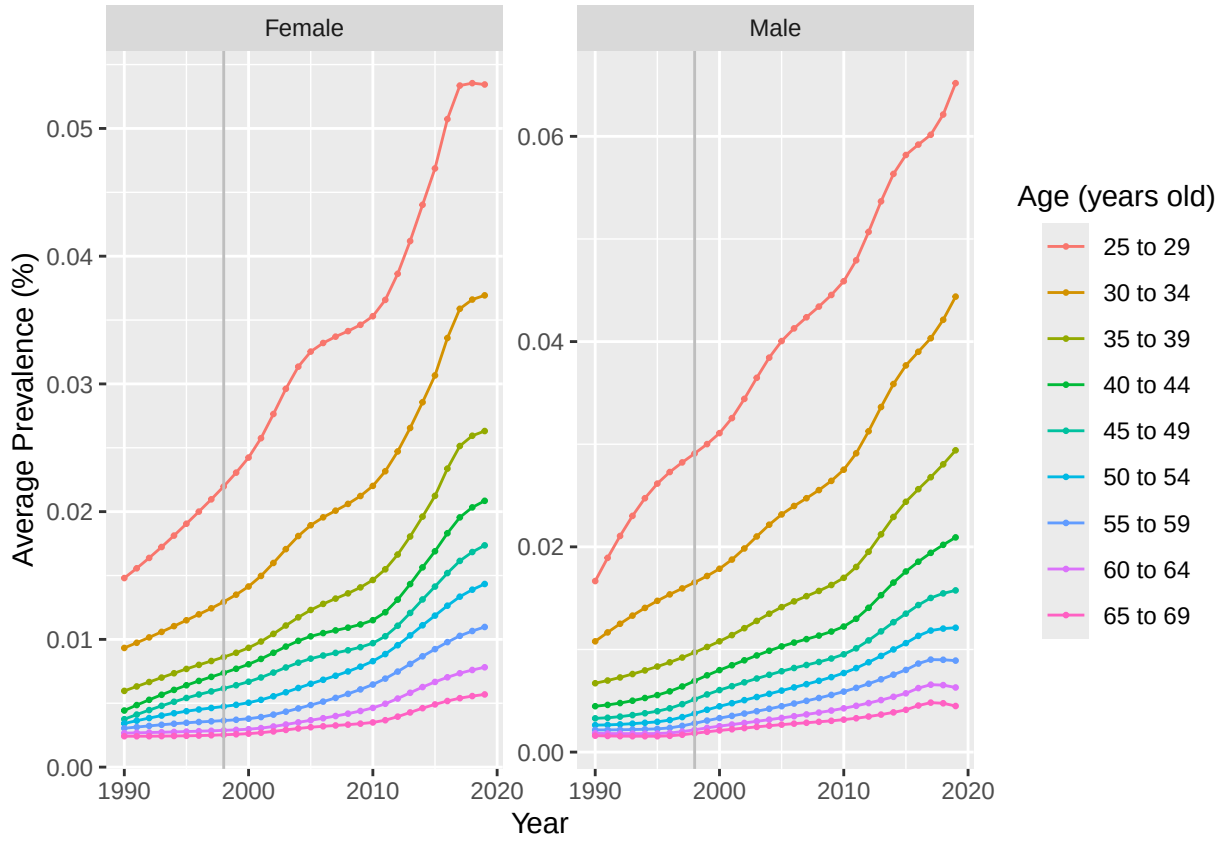



Figure 3.2: The trends of the opioid use related disease prevalence in North America. The plots in this figure are faceted by sex groups, better for visually comparing the prevalence between different age groups.

From Figure 3.1 and Figure 3.2, we can identify that there is an overall increase pattern within each age and sex group. Since the prevalence of the "before" and "after" period are both increasing, and we want to confirm whether the increasing prescription the opioid has an effect on the prevalence, we should compare the speed (slope) of the increase of the prevalence rather than the prevalence itself. If the prevalence of the "after" period is increasing faster than the prevalence of the "before" period, the increasing prescription of the opioid may have an effect.

To test whether the slope of the prevalence of the "before" period is different from the slope of the prevalence of the "after" period, we formulate the hypotheses.

- The null hypothesis (H0): There is no difference in the slope of the prevalence between the "before" period and the "after" period.
- The alternative hypothesis (HA): There is a difference in the slope of the prevalence between the "before" period and the "after" period.

Because the sampling is random, we will use the two-tailed Wilcoxon rank sum test to test the hypotheses.

```
opioid_use_na2 = opioid_use_na %>%
  group_by(sex, age) %>%
  group_split() %>%
  map(~{
    before = diff(.$val[.$year < 1998])
    after = diff(.$val[.$year >= 1998])
    # calculate the slope
```

```
    tibble(
      sex = .$sex[1],
      age = .$age[1],
      p_value = wilcox.test(after, before)$p.value,
      effect_size = median(after) - median(before)
    )
  }) %>%
  bind_rows() %>%
  group_by(sex) %>%
  group_split()
opioid_use_na2
```

```
## <list_of<
##   tbl_df<
##     sex        : character
##     age        : character
##     p_value    : double
##     effect_size: double
##   >
## >[2]>
## [[1]]
## # A tibble: 9 x 4
##   sex    age      p_value effect_size
##   <chr>  <chr>      <dbl>       <dbl>
## 1 Female 25 to 29   0.499    0.000747
## 2 Female 30 to 34   0.249    0.00112
## 3 Female 35 to 39   0.249    0.000811
## 4 Female 40 to 44   0.533    0.000267
## 5 Female 45 to 49   0.405    0.000302
## 6 Female 50 to 54   0.208    0.000487
## 7 Female 55 to 59   0.208    0.000458
## 8 Female 60 to 64   0.208    0.000343
## 9 Female 65 to 69   0.208    0.000195
##
## [[2]]
## # A tibble: 9 x 4
##   sex   age      p_value effect_size
##   <chr> <chr>      <dbl>       <dbl>
## 1 Male  25 to 29   0.678    0.000633
## 2 Male  30 to 34   0.249    0.00131
## 3 Male  35 to 39   0.249    0.00115
## 4 Male  40 to 44   0.249    0.000827
## 5 Male  45 to 49   0.228    0.000729
## 6 Male  50 to 54   0.208    0.000545
## 7 Male  55 to 59   0.155    0.000380
## 8 Male  60 to 64   0.126    0.000292
## 9 Male  65 to 69   0.113    0.000194
```

Table 3.1: The age (age), the sex (sex), the p value (p_value) and the effect size (effect_size) of the the Wilcoxon rank sum test for each group.

From Table 3.1, within each age and sex group, the p value of the Wilcoxon rank sum test is larger than 0.05, we cannot reject the null hypothesis (H0). There is insufficient evidence to conclude that there is a difference in the slope of the prevalence between the "before" period and the "after" period.

However, if we do not consider the sex as a covariate, we will take the average of the prevalence of male and female to represent the prevalence of the whole population within each age group, as the ratio of the male population and the female population within each age group of the divided locations is approximately 1:1.

```
opioid_use_na3 = opioid_use_na %>%
  group_by(year, age) %>%
  summarize(val = mean(val)) %>%
  group_by(age) %>%
  group_split() %>%
  map(~{
    before = diff(.$val[.$year < 1998])
    after = diff(.$val[.$year >= 1998])
    # calculate the slope
    tibble(
      age = .$age[1],
      p_value = wilcox.test(after, before)$p.value,
      effect_size = median(after) - median(before)
    )
  }) %>%
  bind_rows()
opioid_use_na3
```

```
## # A tibble: 9 x 3
##   age         p_value effect_size
##   <chr>         <dbl>       <dbl>
## 1 25 to 29 0.321         0.000182
## 2 30 to 34 0.0000203     0.000486
## 3 35 to 39 0.00000169    0.000323
## 4 40 to 44 0.00385       0.000157
## 5 45 to 49 0.0000507     0.000133
## 6 50 to 54 0.00000169    0.000200
## 7 55 to 59 0.00000338    0.000222
## 8 60 to 64 0.0000760     0.000158
## 9 65 to 69 0.000111      0.000102
```

Table 3.2: The age (age), the p value (p_value) and the effect size (effect_size) of the the Wilcoxon rank sum test for each group.

From Table 3.2, when we do not take the sex as a covariate, we can identify that except 25-29 age group, the p-values for the other age groups are all smaller than 0.05, so we can reject the null hypothesis (H0). Therefore, except 25-29 age group, there is sufficient evidence to conclude that there is a difference in the slope of the prevalence between the "before" period and the "after" period. Additionally, as the effect size (after - before) are all larger than 0, so we can confirm that the increasing prescription of the opioid can accelerate the increase in the prevalence of the opioid use disease except 25-29 age group.

From Figure 3.1 and Figure 3.2, we can assume that there is a linear relationship between the year and the prevalence. We want to use the linear regression model to fit the data, and we test the normality and the homoscedastic of the residuals. To find which age group is the most affected, we compare the difference in the slope of the "before" and "after" period except 25-29 age group.

As the "Multiple R-squared" value in the linear regression model represents the proportion of the variability of the dependent variable explained by the model, will set the threshold of the "Multiple R-squared" as 0.7 (recommended in the ADS2 Lecture 2.6 Slide Page 29).

```
opioid_use_na4 = opioid_use_na %>%
  filter(age != "25 to 29") %>%
  group_by(year, age) %>%
  summarize(val = mean(val)) %>%
  group_by(age) %>%
  group_split() %>%
  map(~{
    before_data = filter(., year < 1998)
    after_data = filter(., year >= 1998)
    before = summary(lm(val ~ year, before_data))
    after = summary(lm(val ~ year, after_data))
    tibble(age = .$age[1],
           before_r_squ = before$r.squared,
           before_slope = before$coefficients[2],
           after_r_squ = after$r.squared,
           after_slope = after$coefficients[2],
           slope_diff = after_slope - before_slope
           )
  }) %>%
  bind_rows() %>%
  arrange(desc(slope_diff))
opioid_use_na4
```

```
## # A tibble: 8 x 6
##   age      before_r_squ before_slope after_r_squ after_slope slope_diff
##   <chr>           <dbl>        <dbl>       <dbl>       <dbl>      <dbl>
## 1 30 to 34        0.999   0.000592         0.962    0.00124    0.000646
## 2 35 to 39        0.999   0.000346         0.946    0.000888   0.000542
## 3 40 to 44        0.997   0.000324         0.937    0.000641   0.000318
## 4 50 to 54        0.994   0.000139         0.968    0.000445   0.000306
## 5 55 to 59        0.974   0.0000632        0.972    0.000345   0.000282
## 6 45 to 49        0.996   0.000252         0.946    0.000516   0.000264
## 7 60 to 64        0.792   0.0000218        0.964    0.000235   0.000213
## 8 65 to 69        0.404   0.00000926       0.948    0.000147   0.000138
```

Table 3.3: The age (age), the "Multiple r-squared" of the "before" period (before_r_squ), the slope of the prevalence in the "before" period (before_slope), the "Multiple r-squared" of the "after" period (after_r_squ), the slope of the prevalence in the "after" period (after_slope), and the difference in the slope between the "before" and "after" period (slope_diff).

From Table 3.3, because the "Multiple r-squared" values are all greater than 0.7 except 65-69 age group (the "before" period), we can compare the difference in the slope between the "before" and "after" period except 65-69 age group. The 30-34 age group has the large difference in the slope of the prevalence of the opioid use disease, so the 30-34 age group is the most affected.

You can use either of the two methods

```
plot(model, 1)
shapiro.test(model$residuals)
```

to further verify the normality the residuals of this model.

You can use

```
plot(model, 2)
```

to further verify whether the residuals of this model are homoscedastic.

Due to the page limit, we do not show the normality and homoscedastic result of the models.

**Conclusion**

If we take the sex as a covariate, we cannot confirm the increasing prescription of the opioid can accelerate the increase in the prevalence of the opioid use disease in each age group. If we do not take the sex as a covariate, we can confirm that the increasing prescription of the opioid can accelerate the increase in the prevalence of the opioid use disease except 25-29 age group, and the 30-34 age group is the most affected.

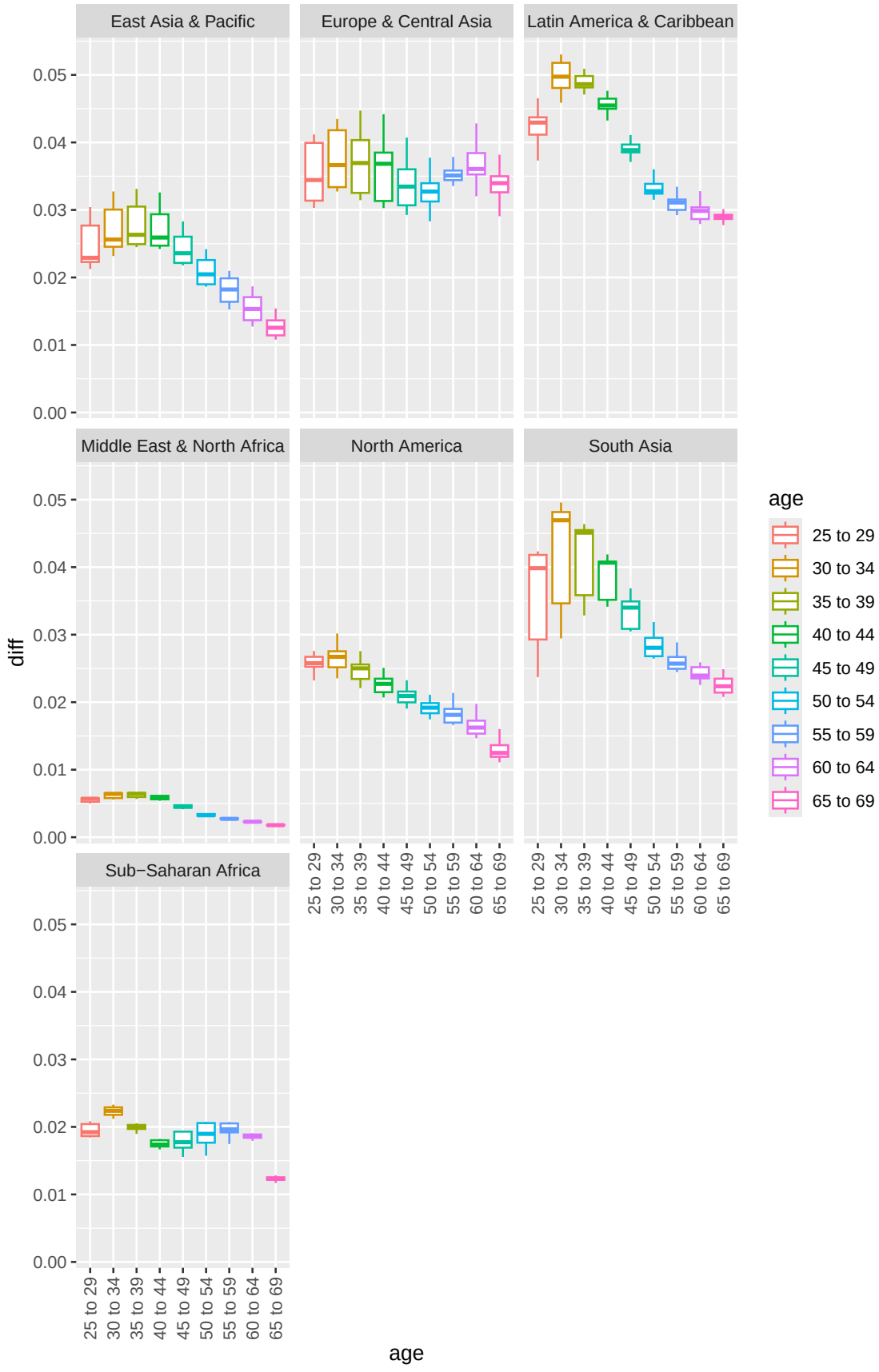
# Part 2: Ask your own question

First, in Figure 2.1, we identify that in east Asia and Pacific, for all age groups, the alcohol-related disease prevalence of the male is higher than that of the female during the years. We are interested in whether in other locations, there are similar patterns. Then, we do some visualization.

```
data1 = substance_use %>%
  filter(measure == "Prevalence",
         cause == "Alcohol use disorders") %>%
  select(location, age, year, sex, val) %>%
  group_by(location, age, year) %>%
  spread(key = "sex", val = "val") %>%
  summarize(diff = Male - Female) %>%
  bind_rows()

data1$location = gsub(data1$location, pattern = " - WB", replacement = "")

ggplot(data1, aes(x = age, y = diff, color = age)) +
  geom_boxplot(outlier.colour = NA,
               outlier.shape = "") +
  facet_wrap(~location) + #, scales = "free_y"
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = .5))
```


East Asia & Pacific
Europe & Central Asia
Latin America & Caribbean
Middle East & North Africa
North America
South Asia
Sub−Saharan Africa
diff
0.00
0.01
0.02
0.03
0.04
0.05
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
age

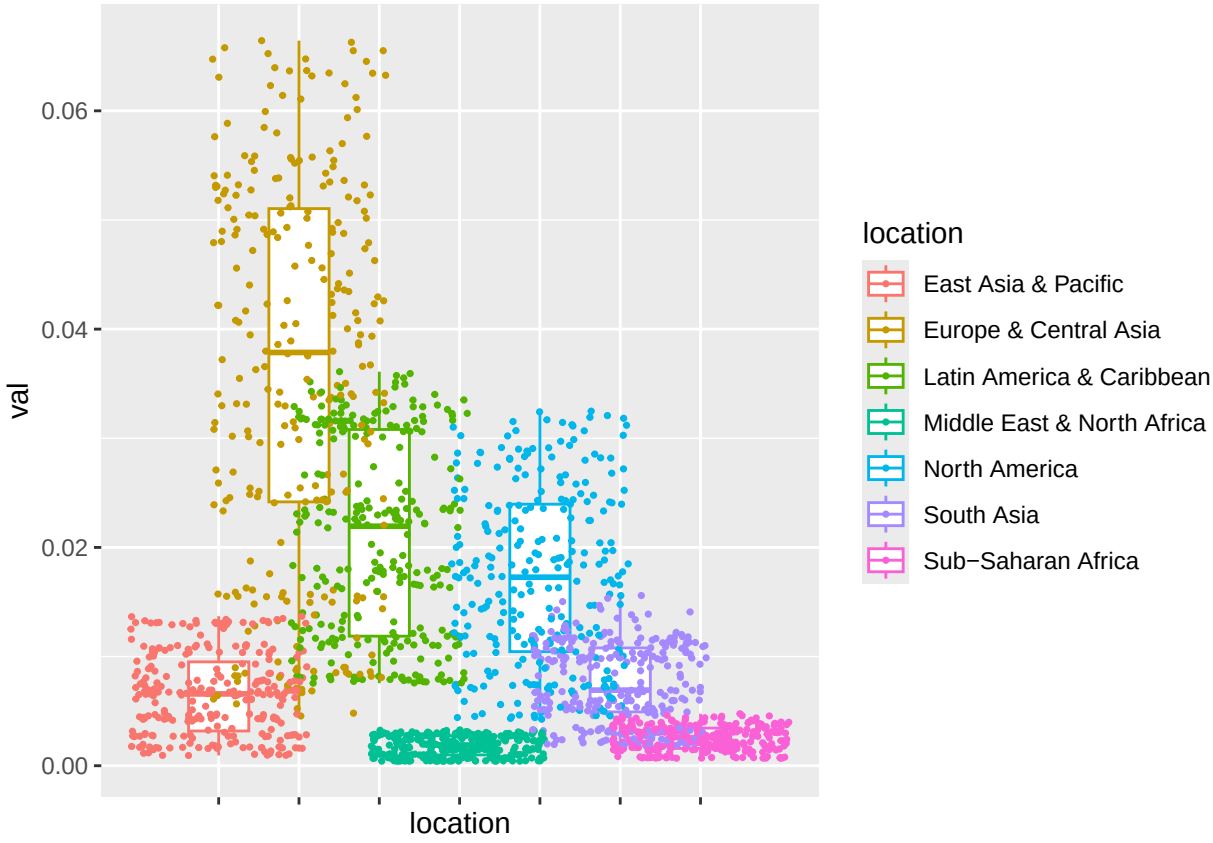
```r
head(substance_use)
```

```
##   measure                 location    sex      age                 cause
## 1  Deaths East Asia & Pacific - WB   Male 25 to 29 Alcohol use disorders
## 2  Deaths East Asia & Pacific - WB Female 25 to 29 Alcohol use disorders
## 3  Deaths East Asia & Pacific - WB   Male 30 to 34 Alcohol use disorders
## 4  Deaths East Asia & Pacific - WB Female 30 to 34 Alcohol use disorders
## 5  Deaths East Asia & Pacific - WB   Male 35 to 39 Alcohol use disorders
## 6  Deaths East Asia & Pacific - WB Female 35 to 39 Alcohol use disorders
##    metric year         val       upper       lower
## 1 Percent 1990 0.004355489 0.005574785 0.003579575
## 2 Percent 1990 0.002316023 0.002622133 0.002052042
## 3 Percent 1990 0.006539015 0.007974114 0.005392593
## 4 Percent 1990 0.002667792 0.002950154 0.002417720
## 5 Percent 1990 0.007597508 0.010585770 0.006359210
## 6 Percent 1990 0.002744876 0.003049935 0.002468063
```

```r
data1 = substance_use %>%
  filter(sex == "Male", measure == "Deaths",
         cause == "Alcohol use disorders") %>%
  #, age == "25 to 29"
  select(location, age, sex, year, val)

data1$location = as.factor(data1$location)
data1$age = as.factor(data1$age)
data1$sex = as.factor(data1$sex)
data1$location = sub(data1$location, pattern = " - WB", replacement = "")

ggplot(data = data1,
       aes(x = location, y = val, color = location)) +
  geom_boxplot() +
  geom_point(position = position_jitterdodge(jitter.width = 2.2),
             size = 0.6) +
  theme(axis.text.x = element_blank())
```

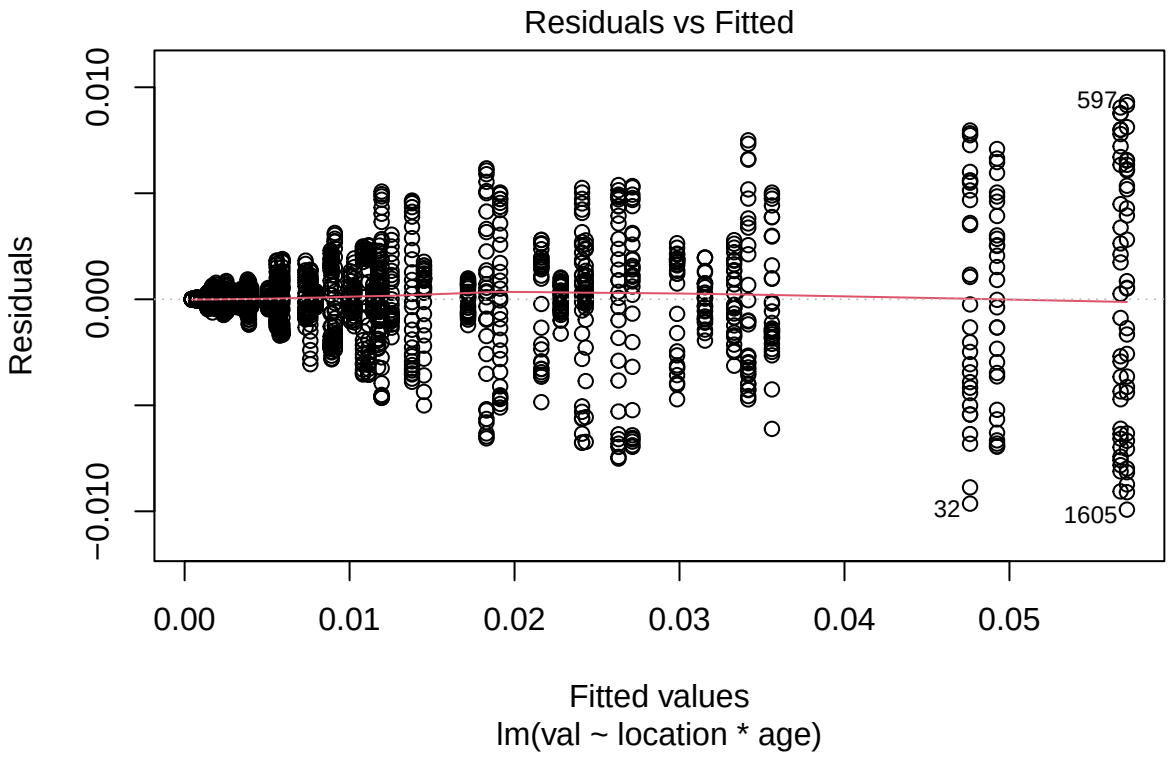



```
model1 = lm(val ~ location * age, data = data1)

shapiro.test(resid(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model1)
## W = 0.88072, p-value < 2.2e-16
```

```
plot(model1, 1)
```




```
summary(model1)
```

```
##
## Call:
## lm(formula = val ~ location * age, data = data1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0099202 -0.0004821  0.0000103  0.0006423  0.0093065
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                    0.0059294  0.0004136  14.337
## locationEurope & Central Asia                  0.0282374  0.0005849  48.279
## locationLatin America & Caribbean              0.0057993  0.0005849   9.915
## locationMiddle East & North Africa            -0.0040706  0.0005849  -6.960
## locationNorth America                          0.0060148  0.0005849  10.284
## locationSouth Asia                             0.0014056  0.0005849   2.403
## locationSub-Saharan Africa                    -0.0042399  0.0005849  -7.249
## age30 to 34                                    0.0029242  0.0005849   5.000
## age35 to 39                                    0.0052200  0.0005849   8.925
## age40 to 44                                    0.0048832  0.0005849   8.349
## age45 to 49                                    0.0029830  0.0005849   5.100
## age50 to 54                                   -0.0001565  0.0005849  -0.268
## age55 to 59                                   -0.0020320  0.0005849  -3.474
```

```
## age60 to 64                                      -0.0035687  0.0005849  -6.102
## age65 to 69                                      -0.0045286  0.0005849  -7.743
## locationEurope & Central Asia:age30 to 34         0.0121577  0.0008271  14.698
## locationLatin America & Caribbean:age30 to 34     0.0069712  0.0008271   8.428
## locationMiddle East & North Africa:age30 to 34 -0.0021182  0.0008271  -2.561
## locationNorth America:age30 to 34                 0.0034301  0.0008271   4.147
## locationSouth Asia:age30 to 34                    0.0022866  0.0008271   2.764
## locationSub-Saharan Africa:age30 to 34           -0.0027029  0.0008271  -3.268
## locationEurope & Central Asia:age35 to 39         0.0177363  0.0008271  21.443
## locationLatin America & Caribbean:age35 to 39     0.0129048  0.0008271  15.602
## locationMiddle East & North Africa:age35 to 39 -0.0041070  0.0008271  -4.965
## locationNorth America:age35 to 39                 0.0069305  0.0008271   8.379
## locationSouth Asia:age35 to 39                   -0.0007449  0.0008271  -0.901
## locationSub-Saharan Africa:age35 to 39           -0.0052743  0.0008271  -6.376
## locationEurope & Central Asia:age40 to 44         0.0176877  0.0008271  21.384
## locationLatin America & Caribbean:age40 to 44     0.0167011  0.0008271  20.191
## locationMiddle East & North Africa:age40 to 44 -0.0038481  0.0008271  -4.652
## locationNorth America:age40 to 44                 0.0103167  0.0008271  12.473
## locationSouth Asia:age40 to 44                   -0.0019322  0.0008271  -2.336
## locationSub-Saharan Africa:age40 to 44           -0.0040460  0.0008271  -4.892
## locationEurope & Central Asia:age45 to 49         0.0104633  0.0008271  12.650
## locationLatin America & Caribbean:age45 to 49     0.0168320  0.0008271  20.350
## locationMiddle East & North Africa:age45 to 49 -0.0024647  0.0008271  -2.980
## locationNorth America:age45 to 49                 0.0113766  0.0008271  13.754
## locationSouth Asia:age45 to 49                   -0.0002642  0.0008271  -0.319
## locationSub-Saharan Africa:age45 to 49           -0.0008495  0.0008271  -1.027
## locationEurope & Central Asia:age50 to 54         0.0015927  0.0008271   1.926
## locationLatin America & Caribbean:age50 to 54     0.0112294  0.0008271  13.576
## locationMiddle East & North Africa:age50 to 54 -0.0002412  0.0008271  -0.292
## locationNorth America:age50 to 54                 0.0073390  0.0008271   8.873
## locationSouth Asia:age50 to 54                   -0.0012829  0.0008271  -1.551
## locationSub-Saharan Africa:age50 to 54            0.0019428  0.0008271   2.349
## locationEurope & Central Asia:age55 to 59        -0.0078204  0.0008271  -9.455
## locationLatin America & Caribbean:age55 to 59     0.0074833  0.0008271   9.047
## locationMiddle East & North Africa:age55 to 59    0.0012071  0.0008271   1.459
## locationNorth America:age55 to 59                 0.0038751  0.0008271   4.685
## locationSouth Asia:age55 to 59                   -0.0002302  0.0008271  -0.278
## locationSub-Saharan Africa:age55 to 59            0.0043782  0.0008271   5.293
## locationEurope & Central Asia:age60 to 64        -0.0160872  0.0008271 -19.449
## locationLatin America & Caribbean:age60 to 64     0.0032767  0.0008271   3.961
## locationMiddle East & North Africa:age60 to 64    0.0024117  0.0008271   2.916
## locationNorth America:age60 to 64                 0.0007312  0.0008271   0.884
## locationSouth Asia:age60 to 64                   -0.0006433  0.0008271  -0.778
## locationSub-Saharan Africa:age60 to 64            0.0049291  0.0008271   5.959
## locationEurope & Central Asia:age65 to 69        -0.0220166  0.0008271 -26.618
## locationLatin America & Caribbean:age65 to 69     0.0007533  0.0008271   0.911
## locationMiddle East & North Africa:age65 to 69    0.0031079  0.0008271   3.757
## locationNorth America:age65 to 69                -0.0018281  0.0008271  -2.210
## locationSouth Asia:age65 to 69                   -0.0006829  0.0008271  -0.826
## locationSub-Saharan Africa:age65 to 69            0.0035950  0.0008271   4.346
##                                                  Pr(>|t|)
## (Intercept)                                       < 2e-16 ***
## locationEurope & Central Asia                     < 2e-16 ***
## locationLatin America & Caribbean                 < 2e-16 ***
```

```
## locationMiddle East & North Africa             4.73e-12 ***
## locationNorth America                           < 2e-16 ***
## locationSouth Asia                             0.016347 *
## locationSub-Saharan Africa                     6.16e-13 ***
## age30 to 34                                    6.29e-07 ***
## age35 to 39                                     < 2e-16 ***
## age40 to 44                                     < 2e-16 ***
## age45 to 49                                    3.74e-07 ***
## age50 to 54                                    0.789064
## age55 to 59                                    0.000524 ***
## age60 to 64                                    1.28e-09 ***
## age65 to 69                                    1.60e-14 ***
## locationEurope & Central Asia:age30 to 34       < 2e-16 ***
## locationLatin America & Caribbean:age30 to 34   < 2e-16 ***
## locationMiddle East & North Africa:age30 to 34 0.010522 *
## locationNorth America:age30 to 34              3.52e-05 ***
## locationSouth Asia:age30 to 34                 0.005760 **
## locationSub-Saharan Africa:age30 to 34         0.001104 **
## locationEurope & Central Asia:age35 to 39       < 2e-16 ***
## locationLatin America & Caribbean:age35 to 39   < 2e-16 ***
## locationMiddle East & North Africa:age35 to 39 7.50e-07 ***
## locationNorth America:age35 to 39               < 2e-16 ***
## locationSouth Asia:age35 to 39                 0.367906
## locationSub-Saharan Africa:age35 to 39         2.29e-10 ***
## locationEurope & Central Asia:age40 to 44       < 2e-16 ***
## locationLatin America & Caribbean:age40 to 44   < 2e-16 ***
## locationMiddle East & North Africa:age40 to 44 3.52e-06 ***
## locationNorth America:age40 to 44               < 2e-16 ***
## locationSouth Asia:age40 to 44                 0.019599 *
## locationSub-Saharan Africa:age40 to 44         1.09e-06 ***
## locationEurope & Central Asia:age45 to 49       < 2e-16 ***
## locationLatin America & Caribbean:age45 to 49   < 2e-16 ***
## locationMiddle East & North Africa:age45 to 49 0.002923 **
## locationNorth America:age45 to 49               < 2e-16 ***
## locationSouth Asia:age45 to 49                 0.749457
## locationSub-Saharan Africa:age45 to 49         0.304546
## locationEurope & Central Asia:age50 to 54      0.054310 .
## locationLatin America & Caribbean:age50 to 54   < 2e-16 ***
## locationMiddle East & North Africa:age50 to 54 0.770611
## locationNorth America:age50 to 54               < 2e-16 ***
## locationSouth Asia:age50 to 54                 0.121076
## locationSub-Saharan Africa:age50 to 54         0.018942 *
## locationEurope & Central Asia:age55 to 59       < 2e-16 ***
## locationLatin America & Caribbean:age55 to 59   < 2e-16 ***
## locationMiddle East & North Africa:age55 to 59 0.144633
## locationNorth America:age55 to 59              3.01e-06 ***
## locationSouth Asia:age55 to 59                 0.780846
## locationSub-Saharan Africa:age55 to 59         1.35e-07 ***
## locationEurope & Central Asia:age60 to 64       < 2e-16 ***
## locationLatin America & Caribbean:age60 to 64  7.74e-05 ***
## locationMiddle East & North Africa:age60 to 64 0.003592 **
## locationNorth America:age60 to 64              0.376833
## locationSouth Asia:age60 to 64                 0.436848
## locationSub-Saharan Africa:age60 to 64         3.03e-09 ***
```

```
## locationEurope & Central Asia:age65 to 69       < 2e-16 ***
## locationLatin America & Caribbean:age65 to 69  0.362567
## locationMiddle East & North Africa:age65 to 69 0.000177 ***
## locationNorth America:age65 to 69              0.027217 *
## locationSouth Asia:age65 to 69                 0.409138
## locationSub-Saharan Africa:age65 to 69         1.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002265 on 1827 degrees of freedom
## Multiple R-squared:  0.9753, Adjusted R-squared:  0.9745
## F-statistic:  1166 on 62 and 1827 DF,  p-value: < 2.2e-16
```

```
# index = c(32, 597, 1605)
#
# data2 = data1[-index, ]
# nrow(data2)
# nrow(data1)
#
# # data2 = data1
# # data2$val = log2(data2$val + 1)
# model2 = aov(val ~ location * age, data = data2)

#shapiro.test(resid(model2))

#plot(model2, 1)

#summary(model)

#TukeyHSD(model)

#levels(data2$age)
#levels(data2$sex)
#model2 = aov(val ~ location + age, data = data2)

#plot(model2, 1)

#anova(model, model2)
```