

Assignment 1

Group 1

2024-11-22

```
library(readr)
library(tidyverse)
library(ggplot2)
library(splines)
```

```
cases <- readr::read_tsv("cases.tsv")
population <- readr::read_tsv("population.tsv")
```

```
##   agegroup      year      sex      n
## Length:1908   Min.   :1970 Length:1908   Min.    : 0.00
## Class :character 1st Qu.:1983 Class :character 1st Qu.:  4.00
## Mode  :character Median :1996 Mode  :character Median : 28.00
##                Mean   :1996                Mean   : 96.73
##                3rd Qu.:2009                3rd Qu.:177.25
##                Max.   :2022                Max.    :621.00
```

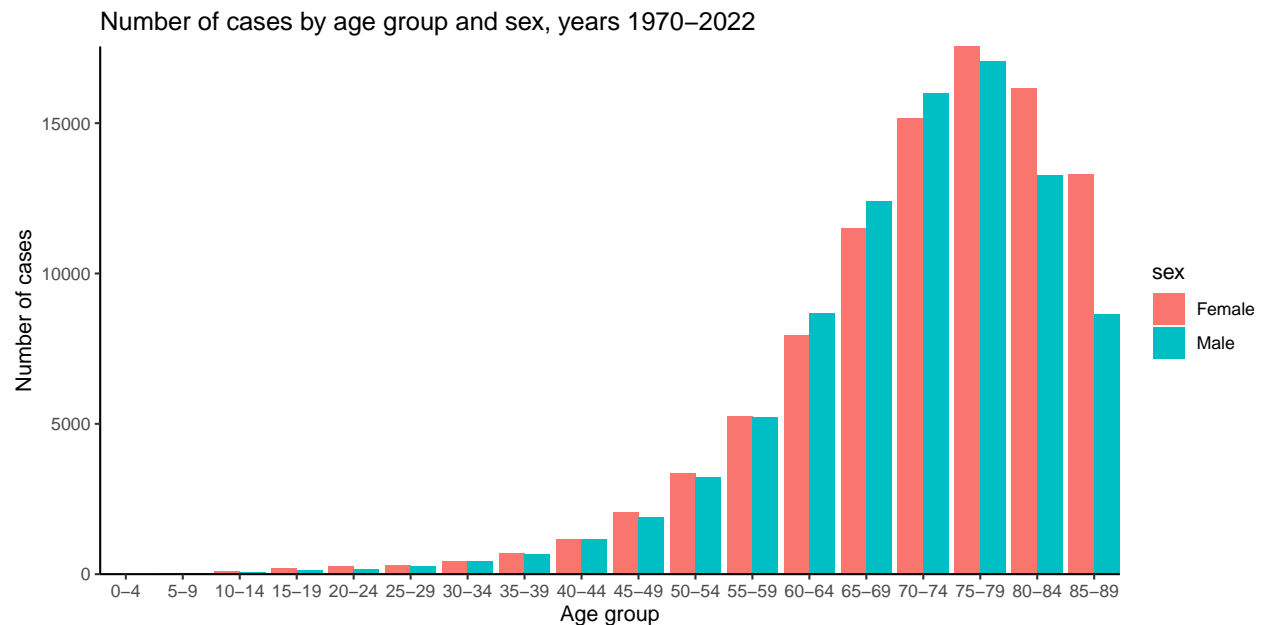
Dataset cases contains the yearly number of reported colon cancer cases (variable “n”) by age group (variable “agegroup”) and sex (variable “sex”), with variable “year” ranging from 1970 to 2022. Eighteen different age groups are defined, as:

```
## [1] 0-4  5-9  10-14 15-19 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59
## [13] 60-64 65-69 70-74 75-79 80-84 85-89
## 18 Levels: 0-4 5-9 10-14 15-19 20-24 25-29 30-34 35-39 40-44 45-49 ... 85-89
```

Two sex groups are defined, as:

```
## [1] "Male" "Female"
```

Question 1: Plot of the number of cases by age group and sex

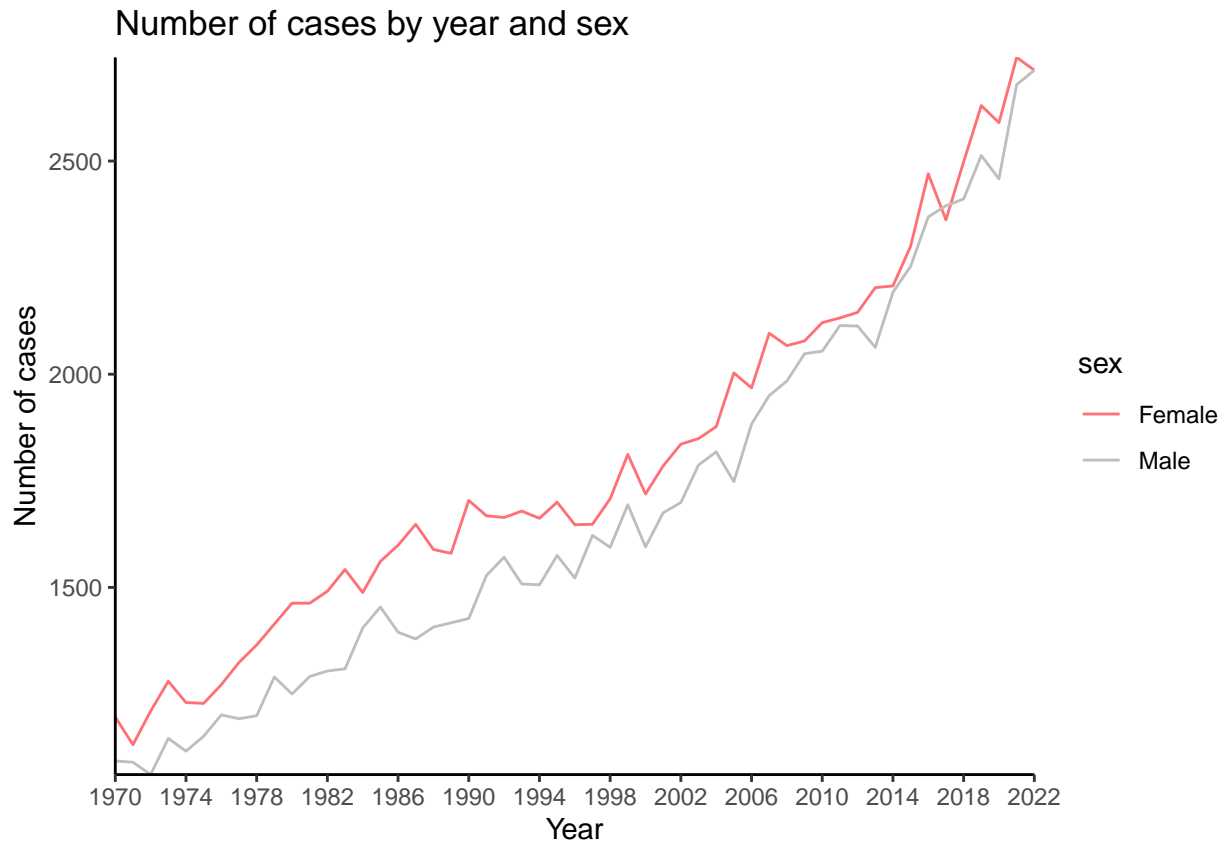


From the plot we can conclude that, both for females and males separately, the total number of reported cases registered between 1970 and 2022 increases with age, reaches a peak at age group 75-79 and then decreases with age.

Question 2: Total number of cases in each calendar year by males and females.

```
cases %>%
  group_by(year, sex) %>%
  summarize(n = sum(n)) %>%
  ungroup() %>%
  ggplot(aes(x = year, y = n, col = sex)) +
  #geom_point() +
  stat_summary(geom="line")+
  scale_color_manual(values=c("#FF7074",
                             "gray"))+

  theme_classic() +
  coord_cartesian(expand = FALSE) +
  scale_x_continuous(lim = c(1970, 2022), breaks = c(seq(1970, 2022, by = 4)))+
  # add labels and title
  ggtitle("Number of cases by year and sex") +
  xlab("Year") + ylab("Number of cases")
```



The graph shows an increasing trend in the yearly number of cases, both for females and males. The total yearly number of reported cancer cases tends to be lower for males than for females.

Question 3: Population

The dataset

The population dataset has the following variables:

```
head(population)
```

```
## # A tibble: 6 x 4
##   agegroup sex    year  n_pop
##   <chr>    <chr> <dbl>  <dbl>
## 1 0-4      Male   1970  296143
## 2 5-9      Male   1970  295092
## 3 10-14    Male   1970  272559
## 4 15-19    Male   1970  281997
## 5 20-24    Male   1970  336648
## 6 25-29    Male   1970  329316
```

```
summary(population)
```

```
##   agegroup          sex          year          n_pop
## Length:1908      Length:1908      Min.    :1970      Min.    : 26513
## Class :character  Class :character  1st Qu.:1983      1st Qu.:215564
## Mode  :character  Mode  :character  Median :1996      Median :269226
##                                     Mean   :1996      Mean   :248347
##                                     3rd Qu.:2009      3rd Qu.:299042
```

```
##                               Max.      :2022   Max.      :400939
```

We can see that it has the same age groups and calendar years as the cases dataset:

```
# Compare age groups
all(unique(cases$agegroup) %in% unique(population$agegroup))
```

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

```
# Compare calendar years
all(unique(cases$year) %in% unique(population$year))
```

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

Heatmap

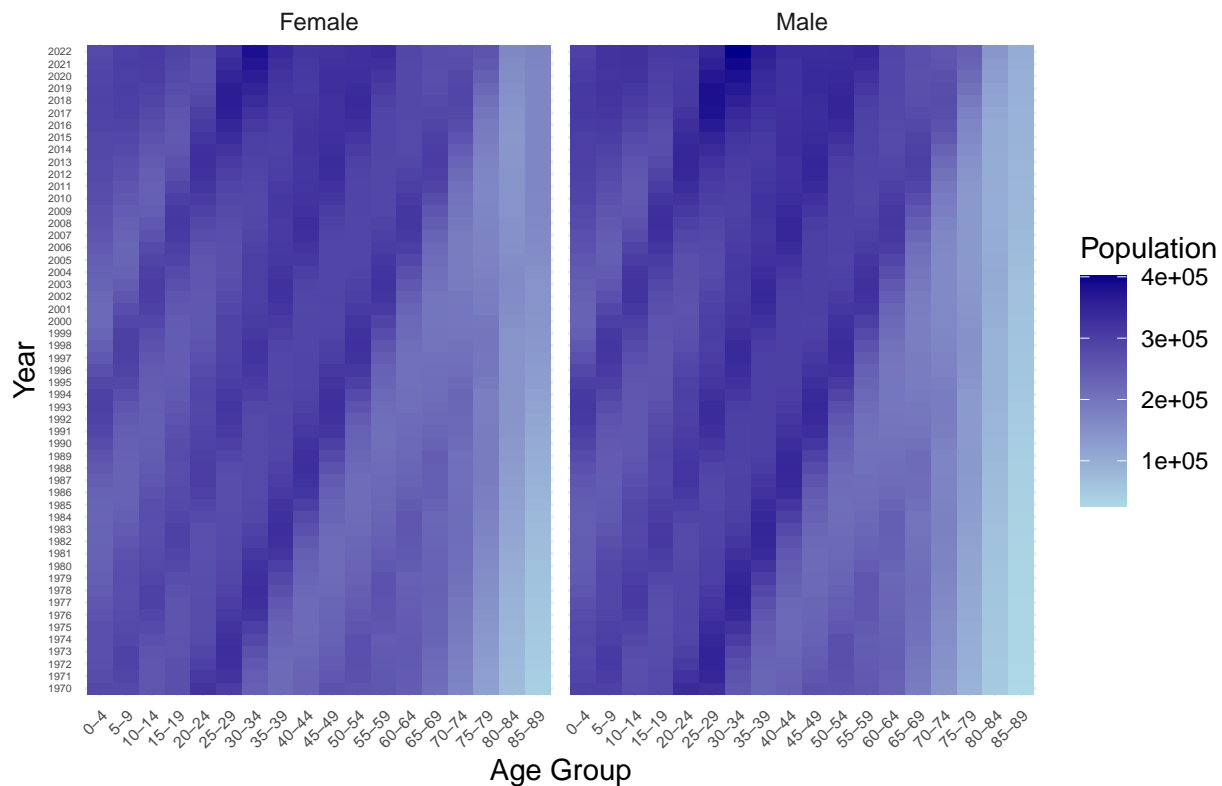
Below we plot an overview of population size by age group and year.

```
data_population <- readr::read_tsv("population.tsv")
data_population$agegroup <- factor(data_population$agegroup, levels = c("0-4", "5-9", "10-14", "15-19",
                                                                      "20-24", "25-29", "30-34", "35-39",
                                                                      "40-44", "45-49", "50-54", "55-59",
                                                                      "60-64", "65-69", "70-74", "75-79",
                                                                      "80-84", "85-89"))

# Summarize total cases by year and sex
population_by_year_sex <- data_population %>%
  group_by(year, sex) %>%
  summarize(total_cases = sum(n_pop, na.rm = TRUE), .groups = "drop")

# Heatmap of population size
ggplot(data_population, aes(x = agegroup, y = factor(year), fill = n_pop)) +
  geom_tile() +
  facet_wrap(~sex) + # Separate plots for males and females
  scale_fill_gradient(low = "lightblue", high = "darkblue", name = "Population") +
  labs(
    title = "Population Size by Age Group and Year",
    x = "Age Group",
    y = "Year"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(size = 6, angle = 45, hjust = 1), # Smaller x-axis text
    axis.text.y = element_text(size = 4) # Smaller y-axis text
  )
```

Population Size by Age Group and Year

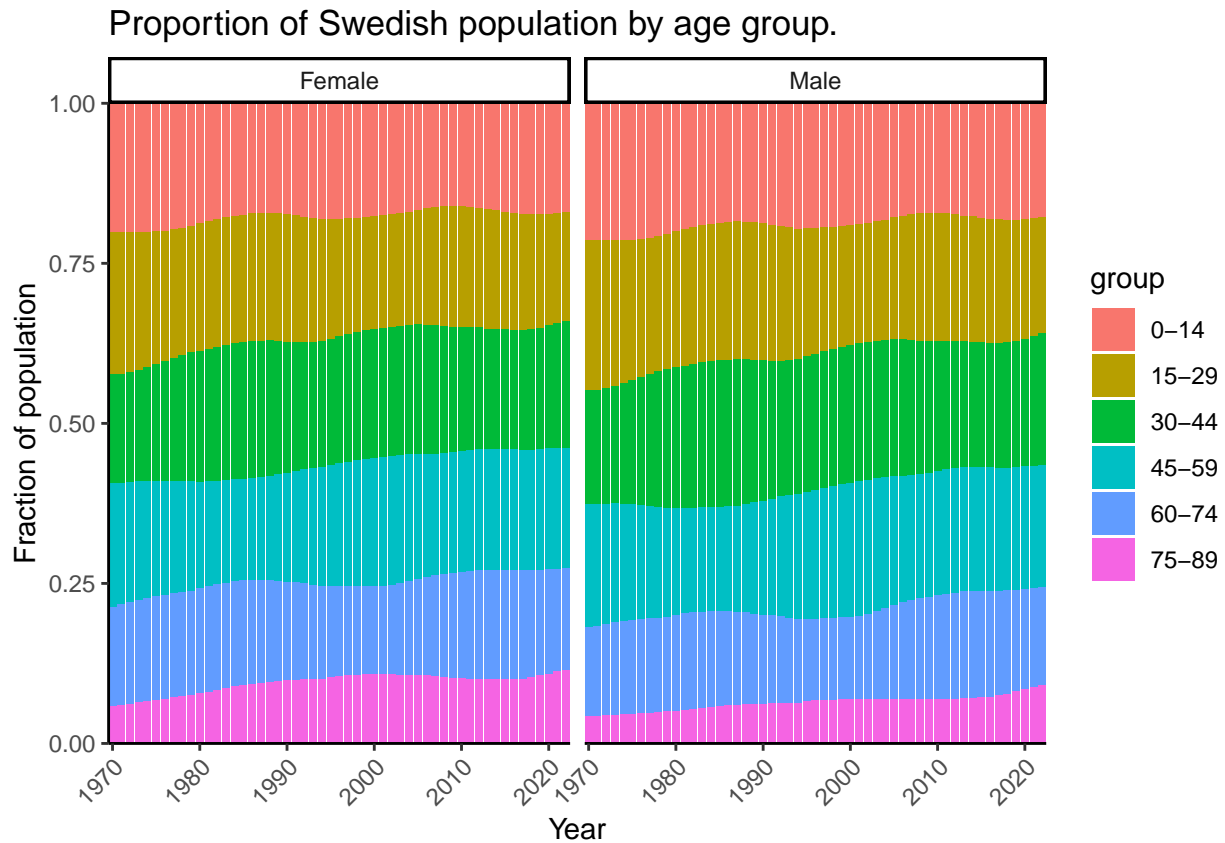


Proportion

The following plot shows the proportions of population in different age groups for each year.

```
population <- population %>% mutate(agegroup = factor(agegroup, levels = factor_groups))
```

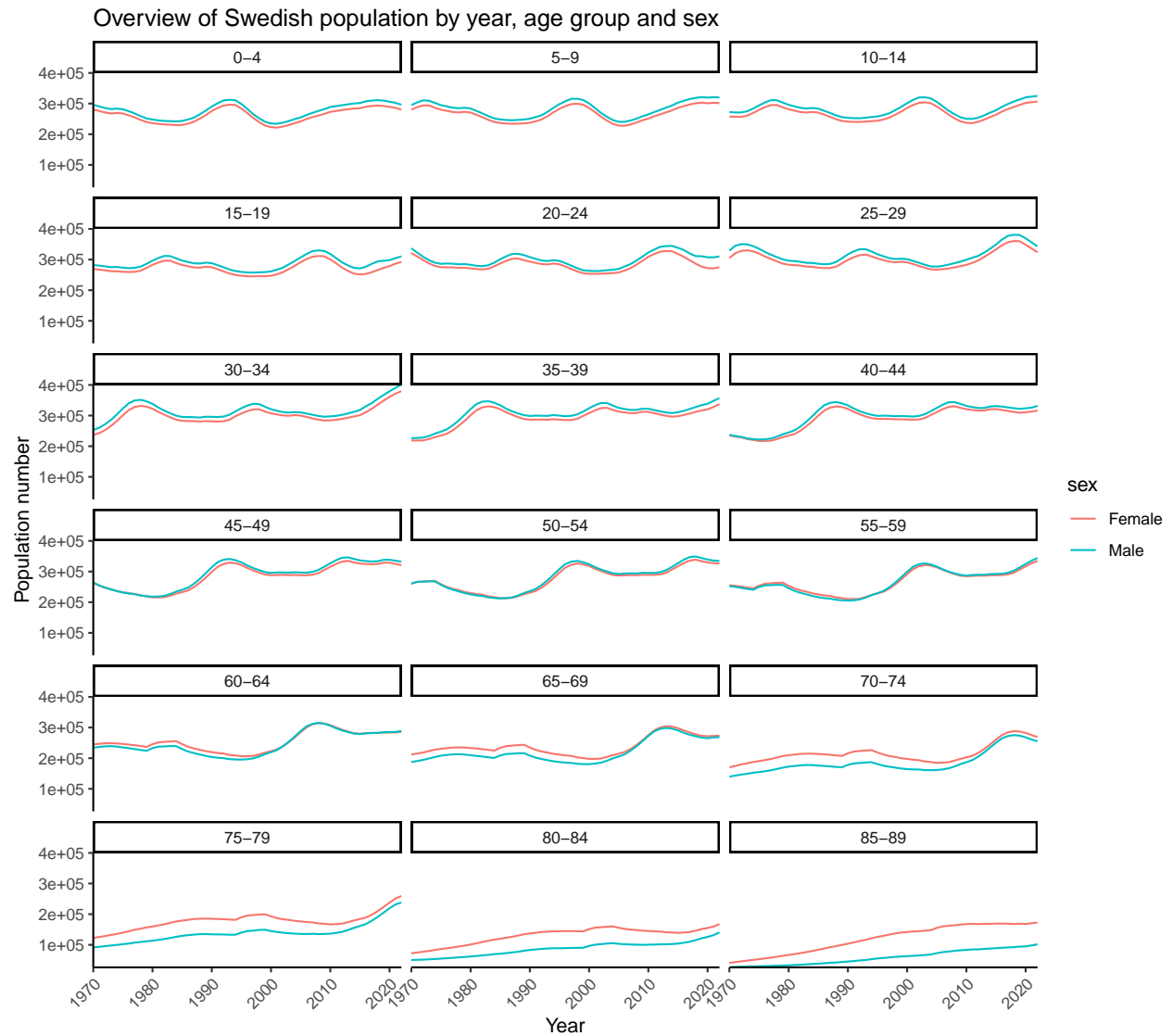
```
population %>%
  mutate(group = case_when(
    agegroup %in% factor_groups[1:3] ~ "0-14",
    agegroup %in% factor_groups[4:6] ~ "15-29",
    agegroup %in% factor_groups[7:9] ~ "30-44",
    agegroup %in% factor_groups[10:12] ~ "45-59",
    agegroup %in% factor_groups[13:15] ~ "60-74",
    agegroup %in% factor_groups[16:18] ~ "75-89",
    TRUE ~ NA
  )) %>%
  ggplot(aes(x = year, y = n_pop, fill = group)) +
  facet_wrap(~sex, ncol = 2) +
  geom_bar(position="fill", stat="identity") +
  theme_classic() +
  theme(axis.text.x = element_text(angle=45, hjust=1)) +
  coord_cartesian(expand = FALSE) +
  # add labels and title
  ggtitle("Proportion of Swedish population by age group.") +
  xlab("Year") + ylab("Fraction of population")
```



Line Plots

The following plots show the population size between genders across different age groups.

```
population %>%
  ggplot(aes(x = year, y = n_pop, color = sex)) +
  facet_wrap(~agegroup, ncol = 3) +
  stat_summary(geom="line")+
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_cartesian(expand = FALSE) +
  # add labels and title
  ggtitle("Overview of Swedish population by year, age group and sex") +
  xlab("Year") + ylab("Population number")
```



Question 4: Merge the datasets

```
# merging datasets
cases_pop <- cases %>% left_join(population)
```

```
cases_pop %>% head()
```

```
## # A tibble: 6 x 5
##   agegroup year sex      n n_pop
##   <fct>    <dbl> <chr> <dbl> <dbl>
## 1 0-4      2022 Male     0 296183
## 2 5-9      2022 Male     0 319820
## 3 10-14    2022 Male     1 325003
## 4 15-19    2022 Male     8 310539
## 5 20-24    2022 Male     5 310354
## 6 25-29    2022 Male     4 342974
```

Cases and population datasets summarized

```
# cases by year and sex
cases_ys <- cases %>%
  group_by(year, sex) %>%
  summarize(n = sum(n)) %>%
  ungroup()

# population by year and sex
population_ys <- population %>%
  group_by(year, sex) %>%
  summarize(n_pop = sum(n_pop)) %>%
  ungroup()

merged <- cases_ys %>%
  left_join(population_ys)
```

Question 5: Incidence rate

```
# rate per ten thousand population, dataset with age
cases_pop <- cases_pop %>%
  # rate per 104 population
  mutate(rate = 104 * n/n_pop)

# rate per ten thousand population, dataset without age
# create dataset without age variable
cases_pop_ys <- merged %>%
  # rate per 104 population
  mutate(rate = 104 * n/n_pop)
```

Interpretation:

Definition of incidence rate: “for a specific follow-up period, the number of new cases of the outcome divided by the total person-time at risk” (Adina Feldman, Lecture 2). In our case the incidence rate is the number of new reported cases of cancer in that year, divided by the total person-time (1 year times population in that year).

In this document, incidence rates are rescaled by a factor of 10^4 to make them more readable, i.e. for each year they correspond to the number of reported cases per 10000 population.

Question 6: Incidence rate by sex and age group

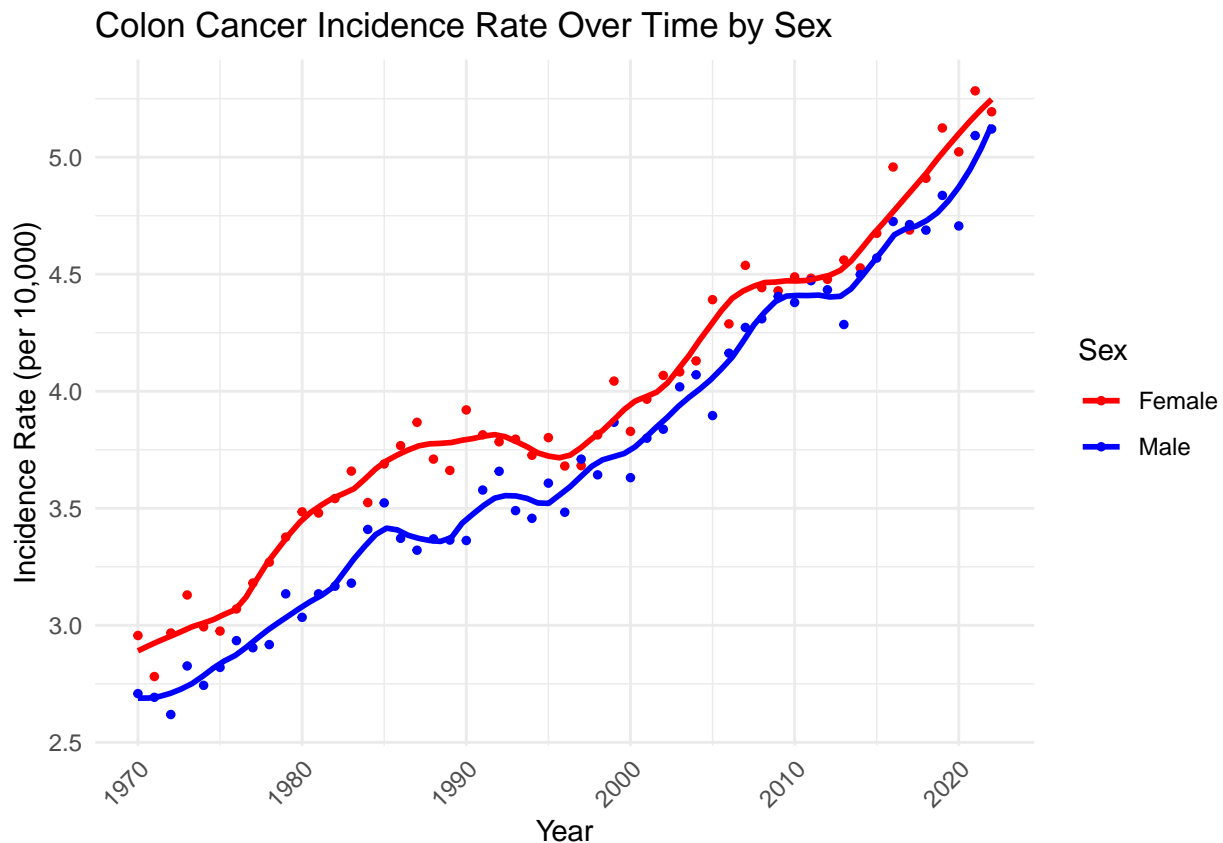
Summarized model

```
# Calculate incidence rate for detailed data
merged_data <- merged_data %>%
  mutate(incidence_rate = n / n_pop * 10000) # Rate per 10000 people

# Calculate incidence rate for summarized data
summary_data <- summary_data %>%
  mutate(incidence_rate = total_cases / total_population * 10000) # Rate per 10000 people
```



```
# Plot incidence rate over time with smoother (separately for males and females)
ggplot(summary_data, aes(x = year, y = incidence_rate, color = sex)) +
  geom_point(size = 1) + # Points for visibility
  geom_smooth(method = "loess", span = 0.2, se = FALSE, linetype = "solid") + # Smoothing with LOESS
  scale_color_manual(values = c("Male" = "blue", "Female" = "red")) +
  labs(
    title = "Colon Cancer Incidence Rate Over Time by Sex",
    x = "Year",
    y = "Incidence Rate (per 10,000)",
    color = "Sex"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The incidence rate per ten thousand population shows an increasing trend in time, both for females and for males separately.

The incidence rate tends to be lower for males than for females.

Including age

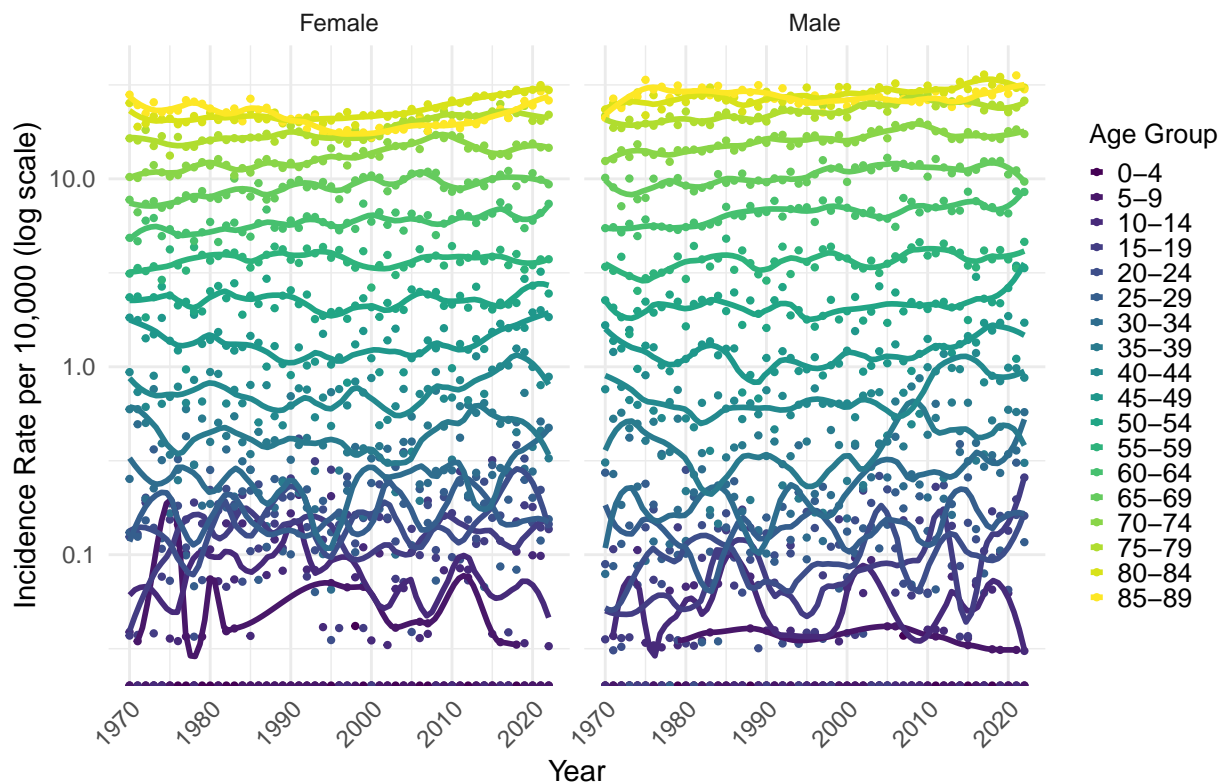
```
# Plot incidence rates over time by sex and age group with smoothed lines
ggplot(merged_data, aes(x = year, y = incidence_rate, color = agegroup, group = agegroup)) +
  geom_point(size = .8) + # Points for visibility
  scale_y_log10() +
  geom_smooth(method = "loess", span = 0.2, se = FALSE, linetype = "solid") + # Smoothing with LOESS
  facet_wrap(~sex) + # Separate plots for males and females
```

```

scale_color_viridis_d(name = "Age Group") + # Use viridis color scale for age groups
labs(
  title = "Colon Cancer Incidence Rate by Sex and Age Group Over Time",
  x = "Year",
  y = "Incidence Rate per 10,000 (log scale)",
  color = "Age Group"
) +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
theme(
  axis.text.x = element_text(angle = 45, hjust = 1),
  legend.key.size = unit(.2, "cm"), # Reduce size of legend keys (symbols)
  legend.text = element_text(size = 10), # Reduce size of legend text
  legend.title = element_text(size = 10) # Adjust size of legend title
)

```

Colon Cancer Incidence Rate by Sex and Age Group Over Time



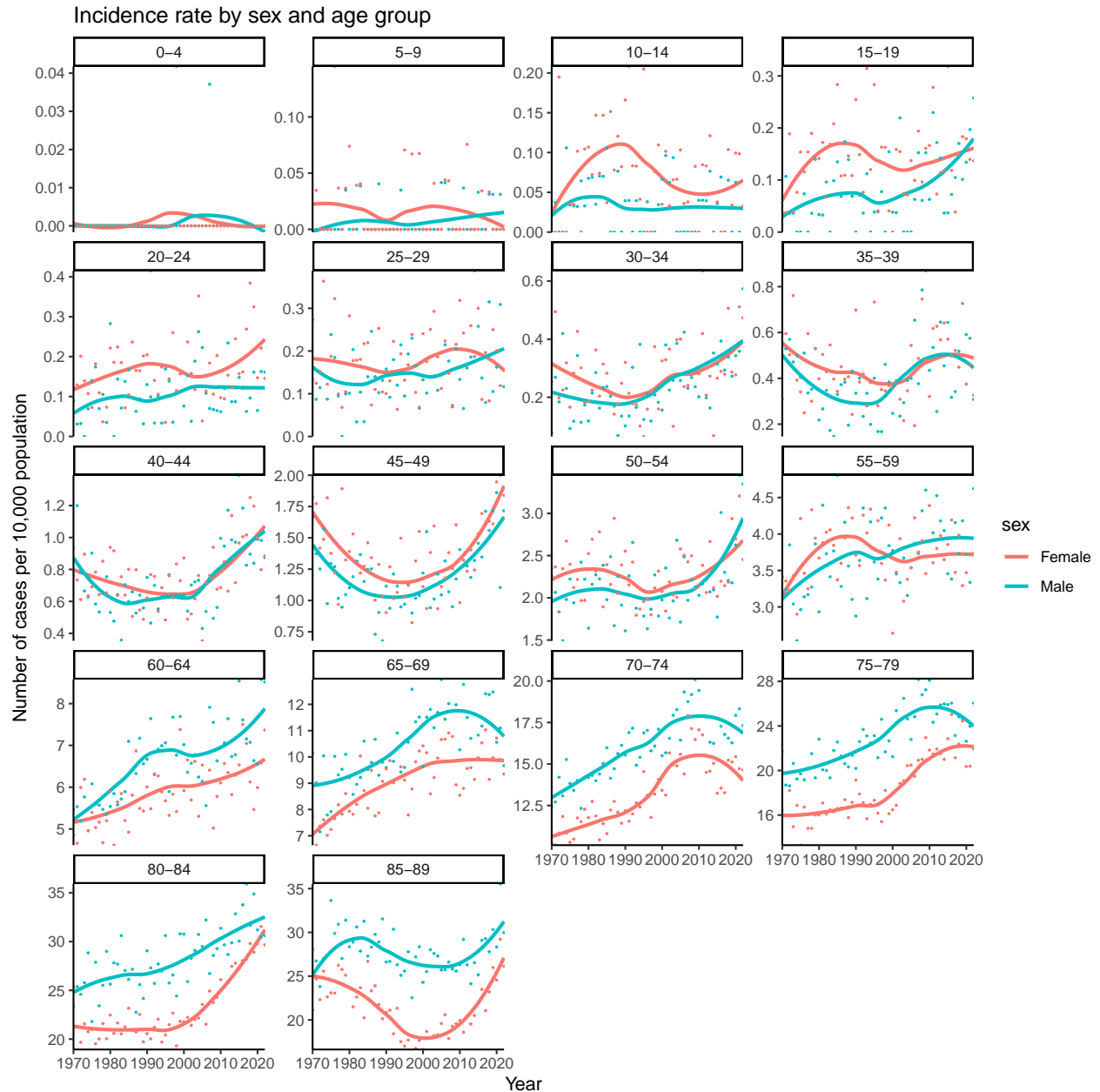
Facet plot with age groups

```

cases_pop %>%
  ggplot(aes(x = year, y = rate, color = sex)) +
  facet_wrap(~ agegroup, ncol = 4, scales = "free_y") +
  # smooth the curve
  geom_smooth(se = FALSE) +
  # non smoothed data
  geom_point(size = 0.2) +
  theme_classic() +

```

```
coord_cartesian(expand = FALSE) +
#scale_x_continuous(lim = c(1970, 2022), breaks = c(seq(1970, 2022, by = 4))) +
# add labels and title
ggtitle(paste("Incidence rate by sex and age group")) +
xlab("Year") + ylab("Number of cases per 10,000 population")
```



From the plots we can conclude:

- Trends for age-specific incidence rates are different for different age groups.
- Comparing females and males, while the crude incidence rate values tend to be higher for females than for males, the age-specific incidence rates have a tendency to be higher for females than for males, when age groups are s.t. age is less than 55. They tend to be lower for females than for males when age groups are s.t. age is greater or equal to 60.

Question 7: Poisson model

```
poisson_model <- cases_pop_ys %>%
  glm(formula = n ~ year + sex + offset(log(n_pop)), family = "poisson")
summary(poisson_model)

##
## Call:
## glm(formula = n ~ year + sex + offset(log(n_pop)), family = "poisson",
##      data = .)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.969e+01  3.071e-01  -96.68  <2e-16 ***
## year         1.094e-02  1.536e-04   71.25  <2e-16 ***
## sexMale      -5.592e-02  4.658e-03  -12.01  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 5505.07  on 105  degrees of freedom
## Residual deviance:  222.81  on 103  degrees of freedom
## AIC: 1211.4
##
## Number of Fisher Scoring iterations: 3
```

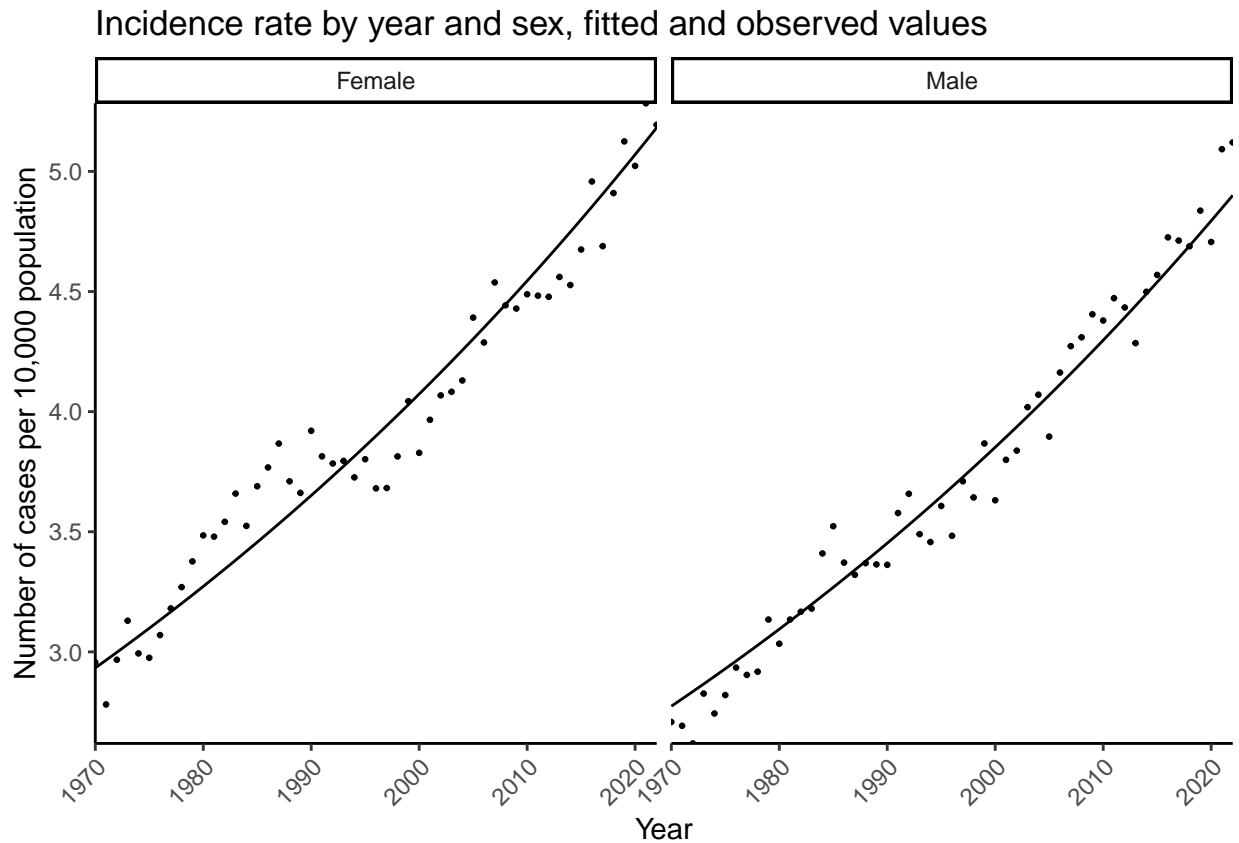
The Poisson model estimates the total number of cases as the dependent variable, using population size as an offset and calendar year and sex as independent variables.

The estimated model is: $\log(n/n_pop) = -29.69415 + 0.01094 \cdot year - 0.05592 \cdot \delta_f(sex)$
with $\delta_f(sex = "Female") = 0$, $\delta_f(sex = "Male") = 1$.

The estimated intercept is -29.69, representing the baseline log-incidence rate when the year is 0 and the reference category for sex (Female). The coefficient for the variable year is 0.01094 ($p < 0.001$), indicating a 1.1% annual increase in the incidence rate ($(e^{(0.01094)} - 1) \cdot 100$), holding sex constant. The coefficient for sex (Male) is -0.05592 ($p < 0.001$), implying a 5.6% lower incidence rate for males compared to females, holding year constant.

```
topplot <- cases_pop_ys %>%
  add_column(fitted = poisson_model$fitted.values) %>%
  mutate(fitted_rate = 10^4 * fitted/n_pop)

topplot %>%
  #pivot_longer(cols = c(rate, fitted_rate), names_to = "type") %>%
  ggplot(aes(x = year)) +
  geom_point(aes(y = rate), size = 0.5) +
  geom_line(aes(y = fitted_rate)) +
  facet_wrap(~sex, ncol = 2) +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_cartesian(expand = FALSE) +
  # add labels and title
  ggtitle("Incidence rate by year and sex, fitted and observed values") +
  xlab("Year") + ylab("Number of cases per 10,000 population")
```



Question 8: Incidence rate values and assumptions

```
# Extract model coefficients
coefficients <- coef(poisson_model)
intercept <- coefficients["(Intercept)"]
year_coeff <- coefficients["year"]
sex_coeff <- coefficients["sexMale"]

# Calculate log incidence rate and incidence rate for 1970 and 2020
incidence_1970_male <- exp(intercept + year_coeff * 1970 + sex_coeff)
incidence_1970_female <- exp(intercept + year_coeff * 1970 )
incidence_2020_male <- exp(intercept + year_coeff * 2020 + sex_coeff)
incidence_2020_female <- exp(intercept + year_coeff * 2020)

# Print results
incidence_1970_male

## (Intercept)
## 0.0002773751
incidence_1970_female

## (Intercept)
## 0.0002933271
incidence_2020_male

## (Intercept)
```

```
## 0.0004794199
incidence_2020_female
```

```
## (Intercept)
## 0.0005069918
```

Based on the model output, the estimated incidence rates in 1970 are 0.0002933 for females and 0.0002774 for males. In 2020, the incidence rates are 0.0005070 for females and 0.0004794 for males.

Assumptions

The assumptions made in this model are:

- The model assumes that the log-incidence rate changes linearly with the calendar year, leading to an annual incidence rate increase of 1.1% ($(e^{(0.01094)} - 1) * 100$).
- A linear relationship between year and log-incidence rate: $\log(\) = _0 + _1(\text{year}) + _2(\text{sex})$
- A constant difference in $\log(\)$ between males and females: $_2 = -0.0559$, males have 5.5% lower incidence rates than females
- No interactions among year, sex, age group.
- Accurate population size as an offset: the population size is accurately measured and representative of the population at risk for both sexes and years.
- Poisson distribution of the dependent variable: the variance in the number of cases increases with the expected number of cases.

Question 9: Adjusting for age

Colon cancer is more common in older age groups, and the age distribution has shifted over time. To better estimate incidence rates, we refit the Poisson model by including interaction terms: “year * sex,” “year * age group,” and “sex * age group.”

```
poisson_model_age <- cases_pop %>%
  mutate(agegroup = factor(agegroup, levels = factor_groups)) %>%
  glm(formula = n ~ year*sex + year*agegroup + agegroup * sex + offset(log(n_pop)), family = "poisson")

# View the summary of the model
summary(poisson_model_age)
```

```
##
## Call:
## glm(formula = n ~ year * sex + year * agegroup + agegroup * sex +
##      offset(log(n_pop)), family = "poisson", data = .)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.879e+01  9.627e+01  -0.715   0.4749
## year           2.618e-02  4.807e-02   0.545   0.5860
## sexMale        1.086e+00  1.545e+00   0.703   0.4819
## agegroup5-9    5.897e+01  9.865e+01   0.598   0.5500
## agegroup10-14  7.020e+01  9.684e+01   0.725   0.4685
## agegroup15-19  3.521e+01  9.654e+01   0.365   0.7153
## agegroup20-24  4.121e+01  9.648e+01   0.427   0.6693
## agegroup25-29  4.711e+01  9.643e+01   0.489   0.6252
## agegroup30-34  3.562e+01  9.637e+01   0.370   0.7117
## agegroup35-39  4.877e+01  9.634e+01   0.506   0.6127
## agegroup40-44  4.329e+01  9.631e+01   0.449   0.6531
```

```

## agegroup45-49      5.194e+01  9.629e+01   0.539   0.5897
## agegroup50-54      5.354e+01  9.628e+01   0.556   0.5782
## agegroup55-59      5.598e+01  9.628e+01   0.581   0.5609
## agegroup60-64      4.976e+01  9.628e+01   0.517   0.6053
## agegroup65-69      4.985e+01  9.627e+01   0.518   0.6046
## agegroup70-74      4.853e+01  9.627e+01   0.504   0.6142
## agegroup75-79      4.823e+01  9.627e+01   0.501   0.6164
## agegroup80-84      4.867e+01  9.627e+01   0.505   0.6132
## agegroup85-89      6.130e+01  9.627e+01   0.637   0.5243
## year:sexMale      -5.696e-04  3.104e-04  -1.835   0.0665
## year:agegroup5-9   -2.792e-02  4.927e-02  -0.567   0.5710
## year:agegroup10-14 -3.283e-02  4.836e-02  -0.679   0.4972
## year:agegroup15-19 -1.496e-02  4.821e-02  -0.310   0.7563
## year:agegroup20-24 -1.788e-02  4.818e-02  -0.371   0.7106
## year:agegroup25-29 -2.080e-02  4.815e-02  -0.432   0.6658
## year:agegroup30-34 -1.484e-02  4.812e-02  -0.308   0.7578
## year:agegroup35-39 -2.117e-02  4.811e-02  -0.440   0.6599
## year:agegroup40-44 -1.816e-02  4.809e-02  -0.378   0.7056
## year:agegroup45-49 -2.220e-02  4.808e-02  -0.462   0.6443
## year:agegroup50-54 -2.274e-02  4.808e-02  -0.473   0.6362
## year:agegroup55-59 -2.372e-02  4.808e-02  -0.493   0.6217
## year:agegroup60-64 -2.037e-02  4.807e-02  -0.424   0.6717
## year:agegroup65-69 -2.020e-02  4.807e-02  -0.420   0.6743
## year:agegroup70-74 -1.935e-02  4.807e-02  -0.403   0.6873
## year:agegroup75-79 -1.904e-02  4.807e-02  -0.396   0.6921
## year:agegroup80-84 -1.914e-02  4.807e-02  -0.398   0.6905
## year:agegroup85-89 -2.552e-02  4.807e-02  -0.531   0.5956
## sexMale:agegroup5-9 -7.823e-01  1.460e+00  -0.536   0.5921
## sexMale:agegroup10-14 -6.568e-01  1.424e+00  -0.461   0.6447
## sexMale:agegroup15-19 -4.856e-01  1.419e+00  -0.342   0.7322
## sexMale:agegroup20-24 -4.125e-01  1.418e+00  -0.291   0.7711
## sexMale:agegroup25-29 -1.100e-01  1.417e+00  -0.078   0.9381
## sexMale:agegroup30-34 -3.506e-02  1.416e+00  -0.025   0.9802
## sexMale:agegroup35-39 -7.101e-02  1.415e+00  -0.050   0.9600
## sexMale:agegroup40-44  1.914e-02  1.415e+00   0.014   0.9892
## sexMale:agegroup45-49 -5.963e-02  1.415e+00  -0.042   0.9664
## sexMale:agegroup50-54 -4.559e-04  1.414e+00   0.000   0.9997
## sexMale:agegroup55-59  4.903e-02  1.414e+00   0.035   0.9723
## sexMale:agegroup60-64  1.673e-01  1.414e+00   0.118   0.9058
## sexMale:agegroup65-69  1.981e-01  1.414e+00   0.140   0.8886
## sexMale:agegroup70-74  2.426e-01  1.414e+00   0.172   0.8638
## sexMale:agegroup75-79  2.643e-01  1.414e+00   0.187   0.8517
## sexMale:agegroup80-84  2.516e-01  1.414e+00   0.178   0.8588
## sexMale:agegroup85-89  3.263e-01  1.414e+00   0.231   0.8175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 406971.1 on 1907 degrees of freedom
## Residual deviance: 2749.5 on 1853 degrees of freedom
## AIC: 11967
##
## Number of Fisher Scoring iterations: 6

```

```

# Extract model coefficients
coefficients <- coef(poisson_model_age)

# males
rate_1970_male <- exp(
  #intercept
  coefficients["(Intercept)"] +
  #year
  coefficients["year"] * 1970 +
  #male
  coefficients["sexMale"] * 1 +
  #age 70-74
  coefficients["agegroup70-74"] * 1 +
  # year*sex
  coefficients["year:sexMale"] * 1970 * 1 +
  # year*age 70-74
  coefficients["year:agegroup70-74"] * 1970 * 1 +
  # sex*age 70-74
  coefficients["sexMale:agegroup70-74"] * 1 * 1)

rate_1970_female <- exp(
  #intercept
  coefficients["(Intercept)"] +
  #year
  coefficients["year"] * 1970 +
  #female
  coefficients["sexMale"] * 0 +
  #age 70-74
  coefficients["agegroup70-74"] * 1 +
  # year*sex
  coefficients["year:sexMale"] * 1970 * 0 +
  # year*age 70-74
  coefficients["year:agegroup70-74"] * 1970 * 1 +
  # sex*age 70-74
  coefficients["sexMale:agegroup70-74"] * 0 * 1)

rate_2020_male <- exp(
  #intercept
  coefficients["(Intercept)"] +
  #year
  coefficients["year"] * 2020 +
  #male
  coefficients["sexMale"] * 1 +
  #age 70-74
  coefficients["agegroup70-74"] * 1 +
  # year*sex
  coefficients["year:sexMale"] * 2020 * 1 +
  # year*age 70-74
  coefficients["year:agegroup70-74"] * 2020 * 1 +
  # sex*age 70-74
  coefficients["sexMale:agegroup70-74"] * 1 * 1)

rate_2020_female <- exp(

```



```

#intercept
coefficients["(Intercept)"] +
#year
coefficients["year"] * 2020 +
#female
coefficients["sexMale"] * 0 +
#age 70-74
coefficients["agegroup70-74"] * 1 +
# year*sex
coefficients["year:sexMale"] * 2020 * 0 +
# year*age 70-74
coefficients["year:agegroup70-74"] * 2020 * 1 +
# sex*age 70-74
coefficients["sexMale:agegroup70-74"] * 0 * 1)

```

```
rate_1970_male
```

```
## (Intercept)
## 0.001349865
```

```
rate_1970_female
```

```
## (Intercept)
## 0.001097564
```

```
rate_2020_male
```

```
## (Intercept)
## 0.001845669
```

```
rate_2020_female
```

```
## (Intercept)
## 0.001544049
```

Based on the updated model output, the estimated incidence rates in 1970 for the 70-74 age group are 0.001350 for males and 0.001098 for females. In 2020, the incidence rates for the same age group are 0.001846 for males and 0.001544 for females.

Plot: fitted values vs observed values

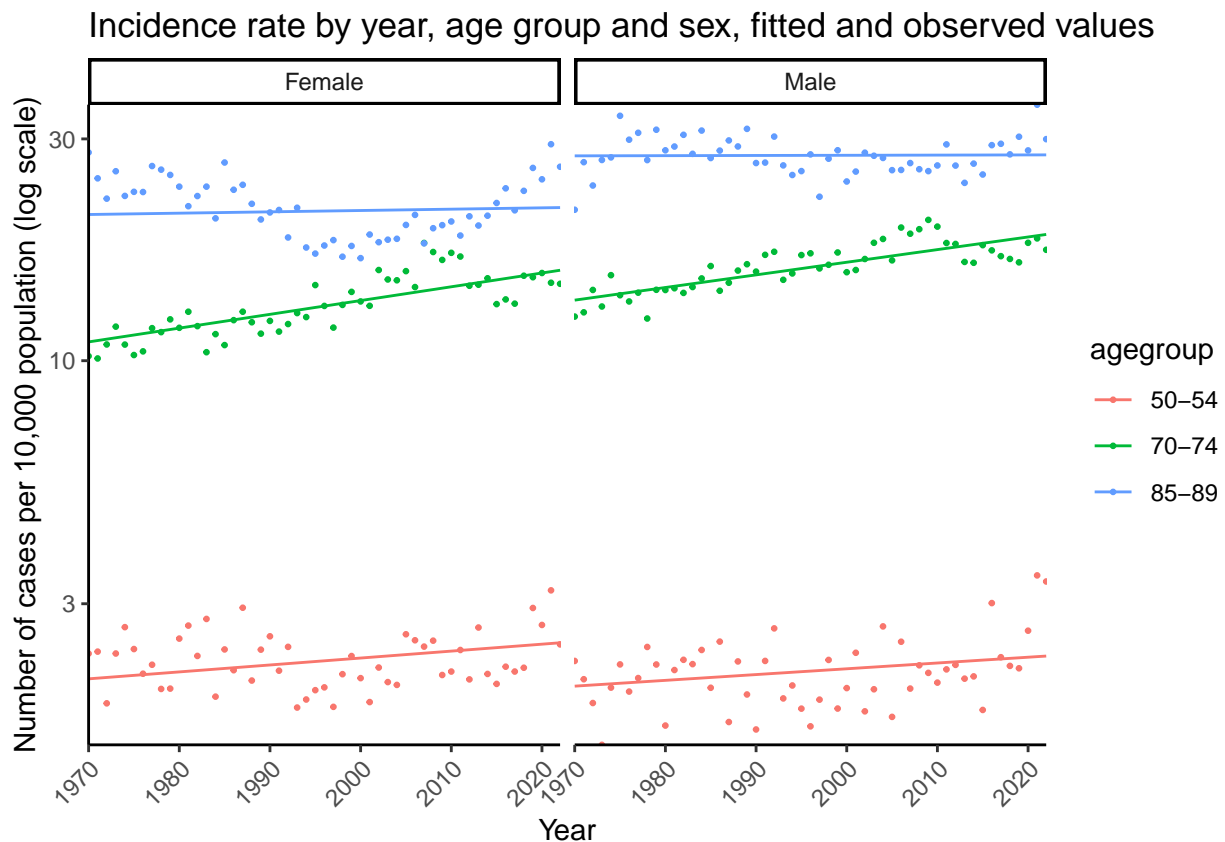
```

toplot <- cases_pop %>%
  add_column(fitted = poisson_model_age$fitted.values) %>%
  mutate(fitted_rate = 10^4 * fitted/n_pop)

toplot %>%
  #pivot_longer(cols = c(rate, fitted_rate), names_to = "type") %>%
  filter(agegroup %in% c("50-54", "70-74", "85-89")) %>%
  ggplot(aes(x = year, color = agegroup)) +
  geom_point(aes(y = rate), size = 0.5) +
  geom_line(aes(y = fitted_rate)) +
  scale_y_log10() +
  facet_wrap(~sex, ncol = 2) +
  theme_classic() +
  theme(axis.text.x = element_text(angle=45, hjust=1)) +
  coord_cartesian(expand = FALSE) +
  #scale_x_continuous(lim = c(1970, 2022), breaks = c(seq(1970, 2022, by = 4))) +

```

```
# add labels and title
ggtitle("Incidence rate by year, age group and sex, fitted and observed values") +
xlab("Year") + ylab("Number of cases per 10,000 population (log scale)")
```



The fit cannot always capture the curvature in the data, for example in the case of females aged 85-89.

Question 10: Splines

```
# Merge cases and population data frames
merged_data <- merge(cases, population, by = c("agegroup", "sex", "year"))
print(head(merged_data, n=5))
```

```
##   agegroup sex year n  n_pop
## 1    0-4 Female 1970 0 280468
## 2    0-4 Female 1971 0 275596
## 3    0-4 Female 1972 0 270765
## 4    0-4 Female 1973 0 267532
## 5    0-4 Female 1974 0 269260
```

```
merged_data <- merged_data %>%
  mutate(
    incidence_rate = (n / n_pop) * 10000,
    age_midpoint = as.numeric(sub("-.*", "", agegroup)) + 2,
    sex = factor(sex)
  )
```

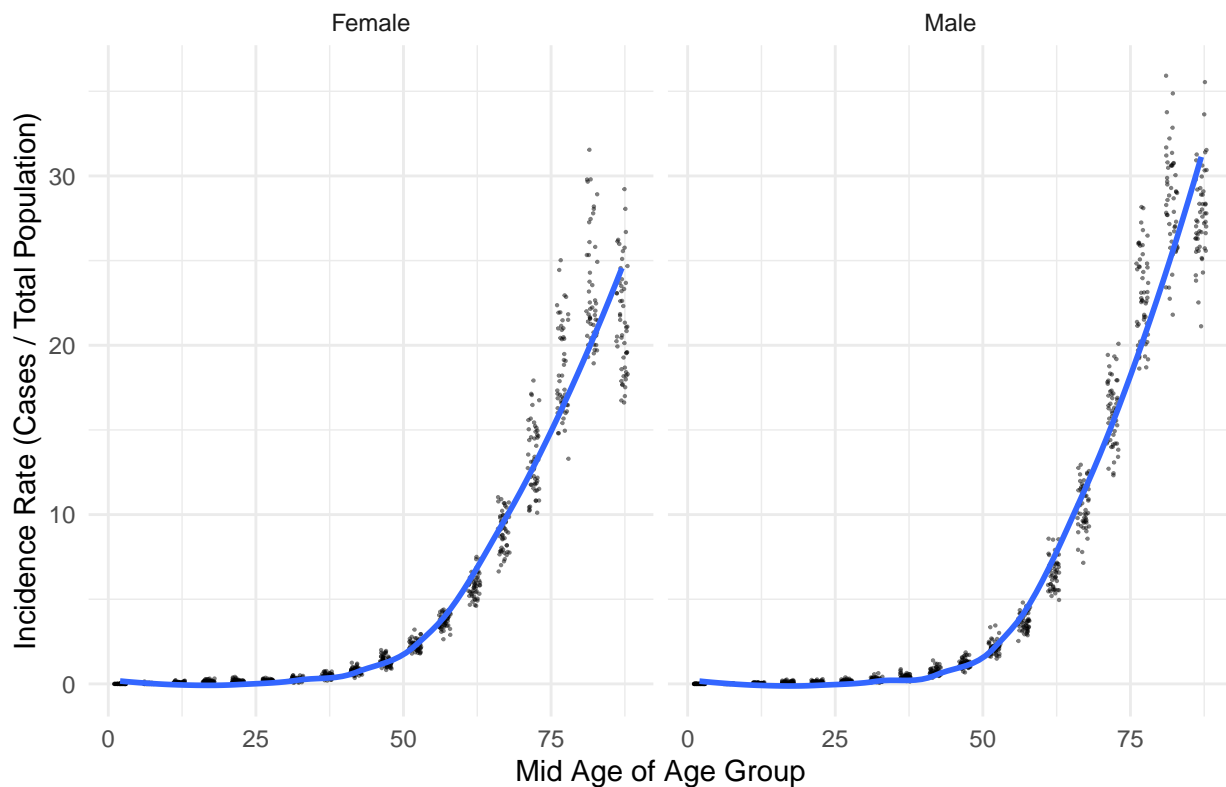
```
# Print column names to verify
```

```
print(colnames(merged_data))
```

```
## [1] "agegroup"      "sex"           "year"          "n"
## [5] "n_pop"        "incidence_rate" "age_midpoint"
```

```
# Plotting the relationship between mid age and incidence rate
ggplot(merged_data, aes(x = age_midpoint, y = incidence_rate)) +
  geom_point(alpha = 0.5, size = .1, position = position_jitter(w = 1, h = 0)) +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ sex) +
  labs(title = "Incidence Rate by Age, with dots corresponding to different calendar years",
       x = "Mid Age of Age Group",
       y = "Incidence Rate (Cases / Total Population)") +
  theme_minimal()
```

Incidence Rate by Age, with dots corresponding to different calendar years



The blue interpolation might be misleading near the endpoints, as one could make the argument that the curve should follow a flatter or even negative curvature at that point. This may just be an effect of the smoothener applied.

```
# Fitting a generalized linear model using splines
splines_model <- glm(
  formula = n ~ bs(year, df = 4)*bs(age_midpoint, df = 4) * sex + offset(log(n_pop)),
  data = merged_data,
  family = poisson()
)

summary(splines_model)
```

```
##
## Call:
## glm(formula = n ~ bs(year, df = 4) * bs(age_midpoint, df = 4) *
##       sex + offset(log(n_pop)), family = poisson(), data = merged_data)
##
## Coefficients:
##
## Estimate Std. Error
## (Intercept) -14.569275 0.748865
## bs(year, df = 4)1 2.501508 1.390446
## bs(year, df = 4)2 2.037596 1.019856
## bs(year, df = 4)3 -0.494946 1.326696
## bs(year, df = 4)4 0.284297 0.956309
## bs(age_midpoint, df = 4)1 2.803894 0.959586
## bs(age_midpoint, df = 4)2 5.501598 0.698437
## bs(age_midpoint, df = 4)3 7.992804 0.775120
## bs(age_midpoint, df = 4)4 8.585834 0.745947
## sexMale -3.346662 1.545262
## bs(year, df = 4)1:bs(age_midpoint, df = 4)1 -2.558816 1.807307
## bs(year, df = 4)2:bs(age_midpoint, df = 4)1 -2.906208 1.344734
## bs(year, df = 4)3:bs(age_midpoint, df = 4)1 2.492452 1.692620
## bs(year, df = 4)4:bs(age_midpoint, df = 4)1 0.560981 1.213426
## bs(year, df = 4)1:bs(age_midpoint, df = 4)2 -2.409901 1.302262
## bs(year, df = 4)2:bs(age_midpoint, df = 4)2 -2.599417 0.959793
## bs(year, df = 4)3:bs(age_midpoint, df = 4)2 -1.518627 1.236974
## bs(year, df = 4)4:bs(age_midpoint, df = 4)2 -0.820693 0.892561
## bs(year, df = 4)1:bs(age_midpoint, df = 4)3 -2.427657 1.440452
## bs(year, df = 4)2:bs(age_midpoint, df = 4)3 -1.326466 1.055886
## bs(year, df = 4)3:bs(age_midpoint, df = 4)3 1.882810 1.368446
## bs(year, df = 4)4:bs(age_midpoint, df = 4)3 0.350943 0.985957
## bs(year, df = 4)1:bs(age_midpoint, df = 4)4 -2.382935 1.385469
## bs(year, df = 4)2:bs(age_midpoint, df = 4)4 -2.578386 1.016811
## bs(year, df = 4)3:bs(age_midpoint, df = 4)4 0.301665 1.322407
## bs(year, df = 4)4:bs(age_midpoint, df = 4)4 -0.119462 0.953222
## bs(year, df = 4)1:sexMale 3.875116 2.704734
## bs(year, df = 4)2:sexMale 0.421600 1.861218
## bs(year, df = 4)3:sexMale 3.127299 2.382851
## bs(year, df = 4)4:sexMale 3.666262 1.731212
## bs(age_midpoint, df = 4)1:sexMale 3.933159 1.879343
## bs(age_midpoint, df = 4)2:sexMale 2.504288 1.444037
## bs(age_midpoint, df = 4)3:sexMale 3.918359 1.586555
## bs(age_midpoint, df = 4)4:sexMale 3.375753 1.538658
## bs(year, df = 4)1:bs(age_midpoint, df = 4)1:sexMale -5.465763 3.346095
## bs(year, df = 4)2:bs(age_midpoint, df = 4)1:sexMale -1.040530 2.353886
## bs(year, df = 4)3:bs(age_midpoint, df = 4)1:sexMale -3.969268 2.936333
## bs(year, df = 4)4:bs(age_midpoint, df = 4)1:sexMale -4.775214 2.122510
## bs(year, df = 4)1:bs(age_midpoint, df = 4)2:sexMale -3.315179 2.533808
## bs(year, df = 4)2:bs(age_midpoint, df = 4)2:sexMale 0.186232 1.747198
## bs(year, df = 4)3:bs(age_midpoint, df = 4)2:sexMale -2.131003 2.225754
## bs(year, df = 4)4:bs(age_midpoint, df = 4)2:sexMale -2.572030 1.618188
## bs(year, df = 4)1:bs(age_midpoint, df = 4)3:sexMale -4.184404 2.781225
## bs(year, df = 4)2:bs(age_midpoint, df = 4)3:sexMale -0.638061 1.915849
## bs(year, df = 4)3:bs(age_midpoint, df = 4)3:sexMale -3.674431 2.446820
## bs(year, df = 4)4:bs(age_midpoint, df = 4)3:sexMale -4.133885 1.777236
## bs(year, df = 4)1:bs(age_midpoint, df = 4)4:sexMale -3.700081 2.693826
```

```

## bs(year, df = 4)2:bs(age_midpoint, df = 4)4:sexMale -0.001042 1.854556
## bs(year, df = 4)3:bs(age_midpoint, df = 4)4:sexMale -2.878592 2.373886
## bs(year, df = 4)4:bs(age_midpoint, df = 4)4:sexMale -3.586707 1.724491
##
## (Intercept) -19.455 < 2e-16 ***
## bs(year, df = 4)1 1.799 0.07201 .
## bs(year, df = 4)2 1.998 0.04572 *
## bs(year, df = 4)3 -0.373 0.70910
## bs(year, df = 4)4 0.297 0.76625
## bs(age_midpoint, df = 4)1 2.922 0.00348 **
## bs(age_midpoint, df = 4)2 7.877 3.35e-15 ***
## bs(age_midpoint, df = 4)3 10.312 < 2e-16 ***
## bs(age_midpoint, df = 4)4 11.510 < 2e-16 ***
## sexMale -2.166 0.03033 *
## bs(year, df = 4)1:bs(age_midpoint, df = 4)1 -1.416 0.15683
## bs(year, df = 4)2:bs(age_midpoint, df = 4)1 -2.161 0.03068 *
## bs(year, df = 4)3:bs(age_midpoint, df = 4)1 1.473 0.14087
## bs(year, df = 4)4:bs(age_midpoint, df = 4)1 0.462 0.64386
## bs(year, df = 4)1:bs(age_midpoint, df = 4)2 -1.851 0.06423 .
## bs(year, df = 4)2:bs(age_midpoint, df = 4)2 -2.708 0.00676 **
## bs(year, df = 4)3:bs(age_midpoint, df = 4)2 -1.228 0.21956
## bs(year, df = 4)4:bs(age_midpoint, df = 4)2 -0.919 0.35784
## bs(year, df = 4)1:bs(age_midpoint, df = 4)3 -1.685 0.09192 .
## bs(year, df = 4)2:bs(age_midpoint, df = 4)3 -1.256 0.20902
## bs(year, df = 4)3:bs(age_midpoint, df = 4)3 1.376 0.16886
## bs(year, df = 4)4:bs(age_midpoint, df = 4)3 0.356 0.72188
## bs(year, df = 4)1:bs(age_midpoint, df = 4)4 -1.720 0.08544 .
## bs(year, df = 4)2:bs(age_midpoint, df = 4)4 -2.536 0.01122 *
## bs(year, df = 4)3:bs(age_midpoint, df = 4)4 0.228 0.81955
## bs(year, df = 4)4:bs(age_midpoint, df = 4)4 -0.125 0.90027
## bs(year, df = 4)1:sexMale 1.433 0.15194
## bs(year, df = 4)2:sexMale 0.227 0.82080
## bs(year, df = 4)3:sexMale 1.312 0.18938
## bs(year, df = 4)4:sexMale 2.118 0.03420 *
## bs(age_midpoint, df = 4)1:sexMale 2.093 0.03636 *
## bs(age_midpoint, df = 4)2:sexMale 1.734 0.08288 .
## bs(age_midpoint, df = 4)3:sexMale 2.470 0.01352 *
## bs(age_midpoint, df = 4)4:sexMale 2.194 0.02824 *
## bs(year, df = 4)1:bs(age_midpoint, df = 4)1:sexMale -1.633 0.10237
## bs(year, df = 4)2:bs(age_midpoint, df = 4)1:sexMale -0.442 0.65845
## bs(year, df = 4)3:bs(age_midpoint, df = 4)1:sexMale -1.352 0.17645
## bs(year, df = 4)4:bs(age_midpoint, df = 4)1:sexMale -2.250 0.02446 *
## bs(year, df = 4)1:bs(age_midpoint, df = 4)2:sexMale -1.308 0.19075
## bs(year, df = 4)2:bs(age_midpoint, df = 4)2:sexMale 0.107 0.91511
## bs(year, df = 4)3:bs(age_midpoint, df = 4)2:sexMale -0.957 0.33835
## bs(year, df = 4)4:bs(age_midpoint, df = 4)2:sexMale -1.589 0.11196
## bs(year, df = 4)1:bs(age_midpoint, df = 4)3:sexMale -1.505 0.13245
## bs(year, df = 4)2:bs(age_midpoint, df = 4)3:sexMale -0.333 0.73910
## bs(year, df = 4)3:bs(age_midpoint, df = 4)3:sexMale -1.502 0.13317
## bs(year, df = 4)4:bs(age_midpoint, df = 4)3:sexMale -2.326 0.02002 *
## bs(year, df = 4)1:bs(age_midpoint, df = 4)4:sexMale -1.374 0.16958
## bs(year, df = 4)2:bs(age_midpoint, df = 4)4:sexMale -0.001 0.99955
## bs(year, df = 4)3:bs(age_midpoint, df = 4)4:sexMale -1.213 0.22528
## bs(year, df = 4)4:bs(age_midpoint, df = 4)4:sexMale -2.080 0.03754 *

```

```

## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 406971.1  on 1907  degrees of freedom
## Residual deviance:  2572.3   on 1858  degrees of freedom
## AIC: 11780
##
## Number of Fisher Scoring iterations: 6

# Assuming you have your merged data as 'merged_data' and fitted model as 'model'
# Create a new data frame for predictions
age_groups <- c(52, 72, 87)
sex_groups <- c("Female", "Male")
years <- unique(merged_data$year) # Get all unique years from your data

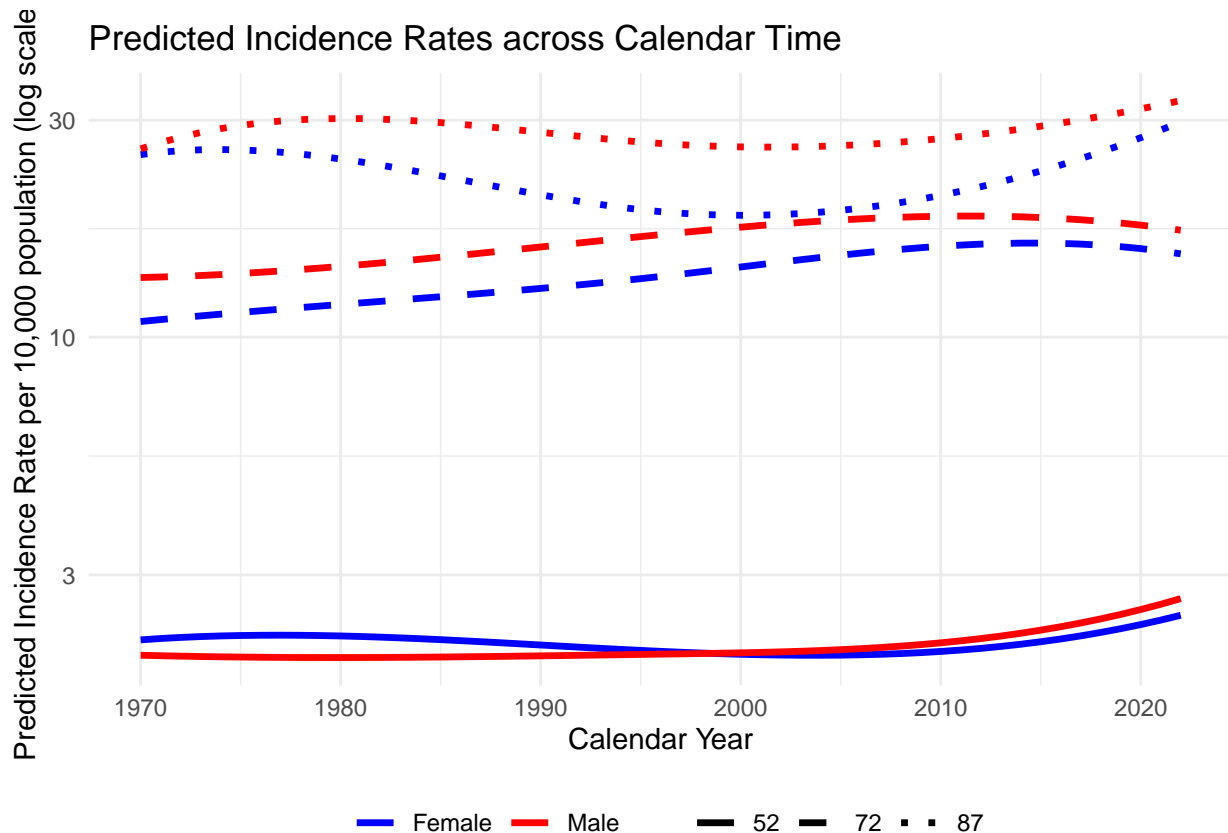
# Create an empty data frame to store predictions
prediction_data <- expand.grid(
  year = years,
  age_midpoint = age_groups,
  sex = sex_groups
)

# Add population to avoid warnings (use dummy values as predictions will standardize population)
prediction_data$n_pop <- 1

# Use the predict() function to generate predicted values
prediction_data$predicted_incidence <- predict(splines_model, newdata = prediction_data, type = "response")

prediction_data %>%
  mutate(predicted_incidence = 10^4 * predicted_incidence) %>%
  ggplot(aes(x = year, y = predicted_incidence, color = sex, linetype = as.factor(age_midpoint))) +
    geom_line(size = 1.2) +
    scale_y_log10() +
    labs(title = "Predicted Incidence Rates across Calendar Time",
         x = "Calendar Year",
         y = "Predicted Incidence Rate per 10,000 population (log scale)") +
    scale_color_manual(values = c("blue", "red")) +
    scale_linetype_manual(values = c("solid", "dashed", "dotted")) +
    theme_minimal() +
    theme(legend.title = element_blank()) +
    theme(legend.position = "bottom")

```

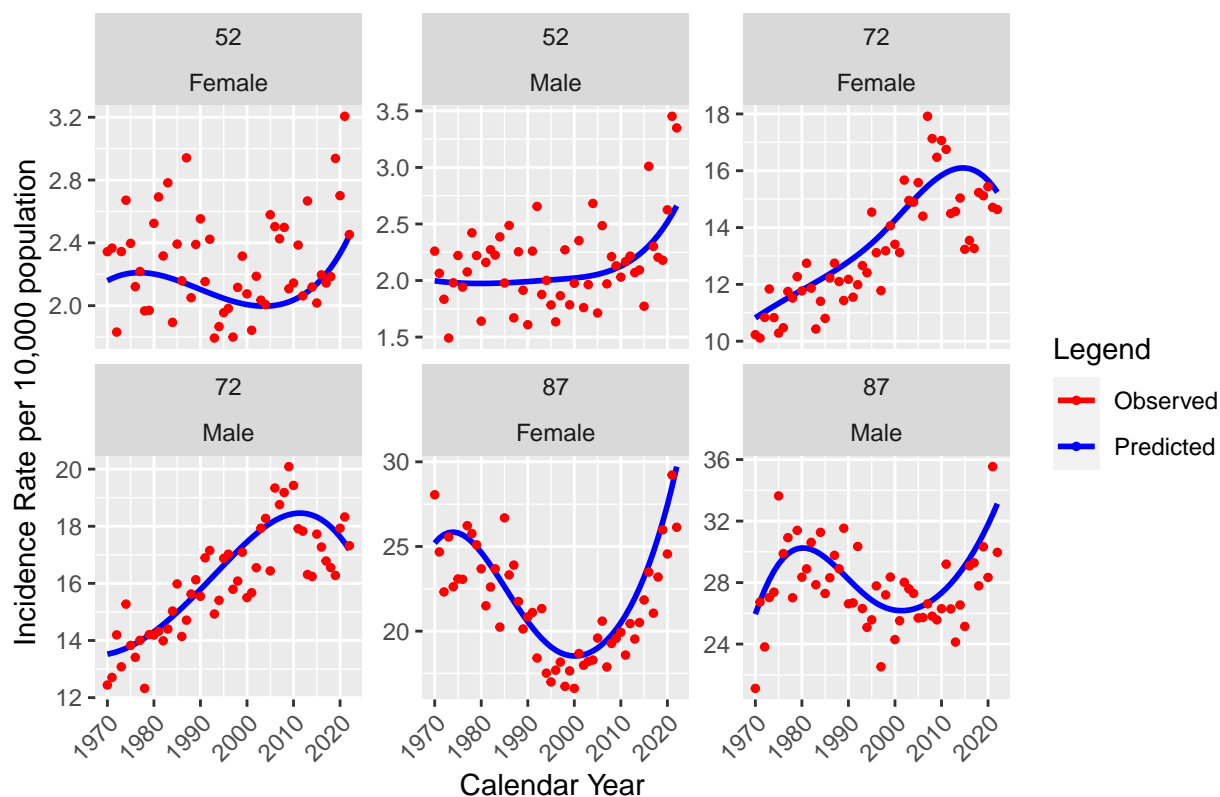


```
# Compare with observed values
observed_data <- merged_data %>%
  filter(age_midpoint %in% c(52, 72, 87)) %>%
  group_by(year, age_midpoint, sex) %>%
  summarize(observed_rate = sum(n) / sum(n_pop))

# Combine predicted and observed data
comparison_data <- left_join(prediction_data, observed_data, by = c("year", "age_midpoint", "sex"))

comparison_data %>%
  mutate(predicted_incidence = 10^4 * predicted_incidence,
         observed_rate = 10^4 * observed_rate) %>%
  ggplot(aes(x = year)) +
  geom_line(aes(y = predicted_incidence, color = "Predicted"), size = 1) +
  geom_point(aes(y = observed_rate, color = "Observed"), size = 1) +
  facet_wrap(~ age_midpoint + sex, scales = "free_y") +
  labs(
    title = "Comparison of Predicted and Observed Incidence Rates",
    x = "Calendar Year",
    y = "Incidence Rate per 10,000 population",
    color = "Legend"
  ) +
  scale_color_manual(values = c("Predicted" = "blue", "Observed" = "red")) +
  theme(axis.text.x = element_text(angle=45, hjust=1))
```

Comparison of Predicted and Observed Incidence Rates



Compared to the fit without splines, the plot shows that the fitted curves follow relatively well the distribution of the data.

The computed AIC is 11780 and it is lower than the AIC of the fit without splines (11967), indicating a better fit.

Question 11: Direct age standardised rate

```
# reference values for male and female population
standard_pop <- population %>%
  filter(year == 2022) %>%
  rename(ref_pop = n_pop) %>%
  select(-year)

# reference value for total male population
tot_male <- standard_pop %>%
  filter(sex == "Male") %>%
  pull(ref_pop) %>% sum()

# reference value for total female population
tot_female <- standard_pop %>%
  filter(sex == "Female") %>%
  pull(ref_pop) %>% sum()
```

Standardised incidence rates


```

age_standardised_rates <- cases_pop %>%
  select(-rate) %>%
  left_join(standard_pop) %>%
  # add number of cases obtained if population number was ref_pop
  mutate(cases_ref_pop = ref_pop * n/n_pop) %>%
  # for each year, compute total (sum over agegroups) and divide by total reference population
  # keep male and female categories
  group_by(year, sex) %>%
  # sum over agegroups
  summarize(sum_ages = sum(cases_ref_pop, na.rm = TRUE)) %>%
  mutate(standardised_rate = case_when(sex == "Female" ~ sum_ages/tot_female,
                                       sex == "Male" ~ sum_ages/tot_male,
                                       TRUE ~ NA)
        ) %>%
  select(-sum_ages) %>%
  # rescale
  mutate(standardised_rate = 10^4 * standardised_rate)

```

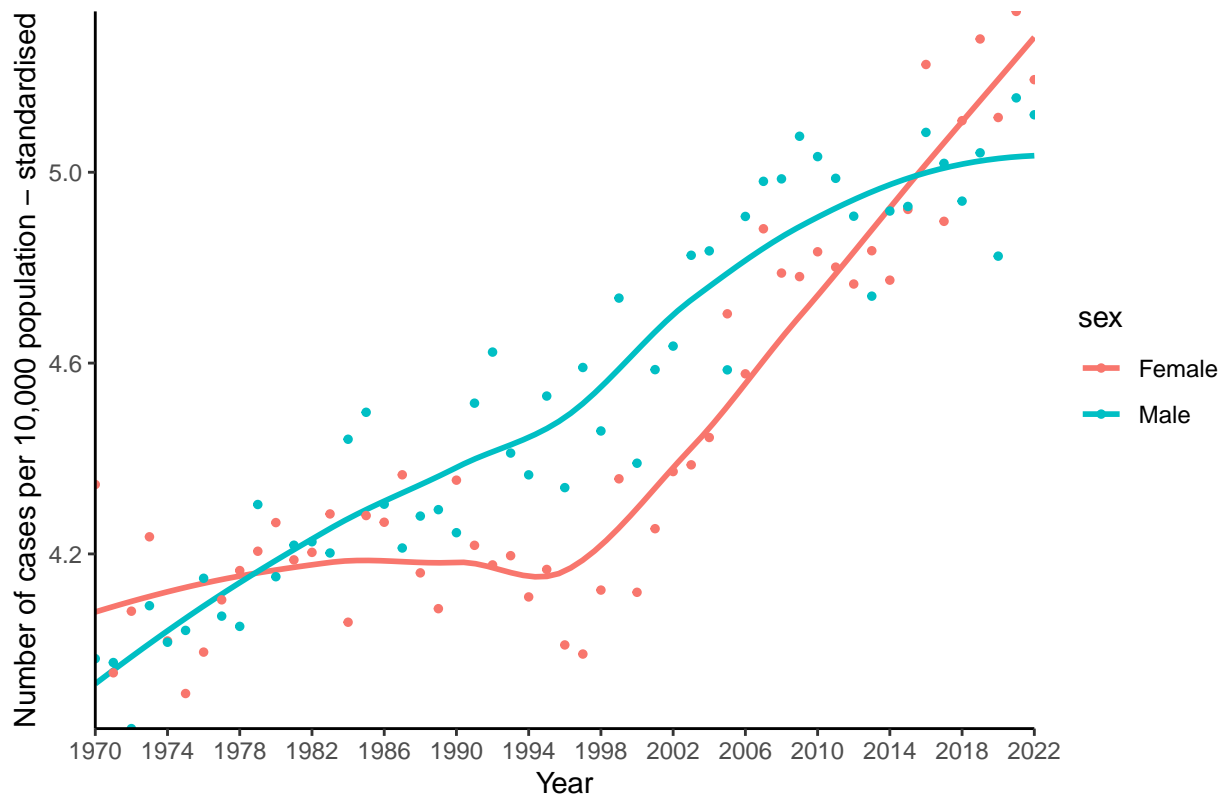
Plot of the age standardised rates

```

age_standardised_rates %>%
  ggplot(aes(x = year, y = standardised_rate, color = sex)) +
  # smooth the curve
  geom_smooth(se = FALSE) +
  # non smoothed data
  geom_point(size = 1) +
  theme_classic() +
  coord_cartesian(expand = FALSE) +
  scale_x_continuous(lim = c(1970, 2022), breaks = c(seq(1970, 2022, by = 4))) +
  # add labels and title
  ggtitle("Standardised incidence rate: cases per 10^4 population by year and sex") +
  xlab("Year") + ylab("Number of cases per 10,000 population - standardised")

```

Standardised incidence rate: cases per 10⁴ population by year and sex

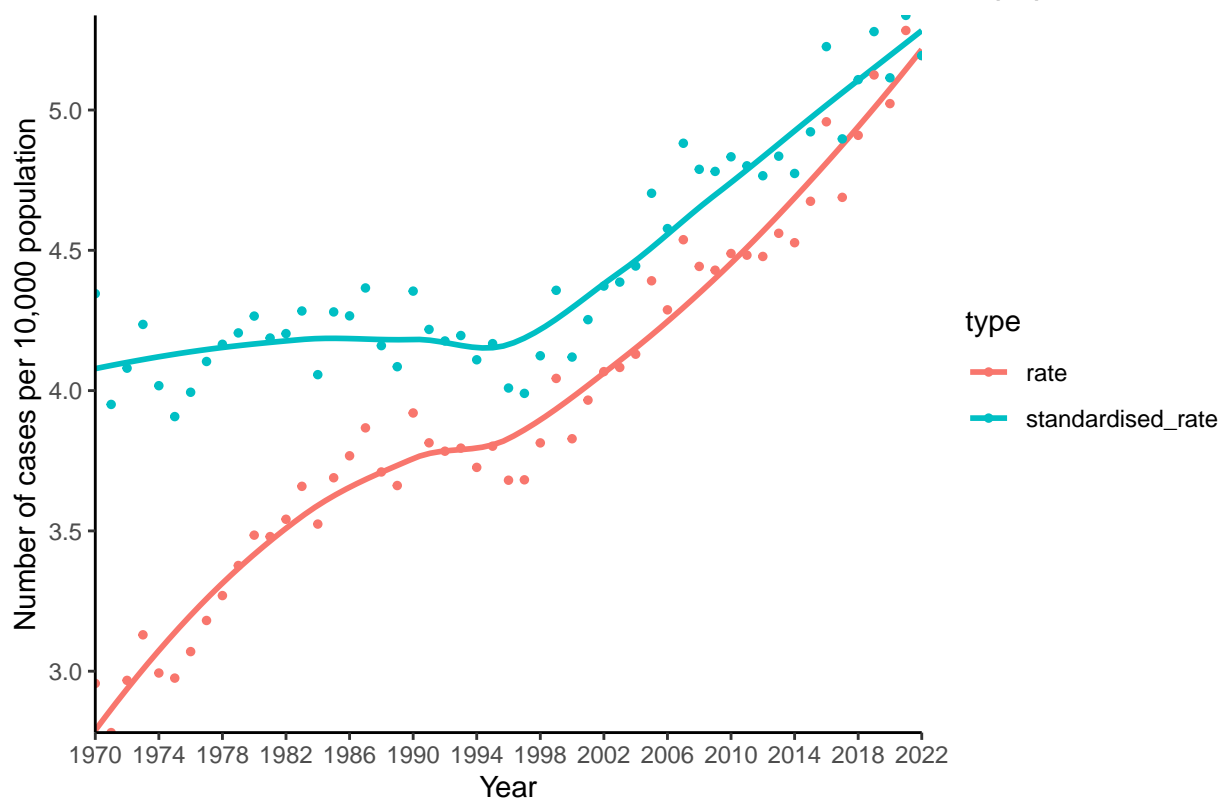


Comparison with non standardised rates

Females

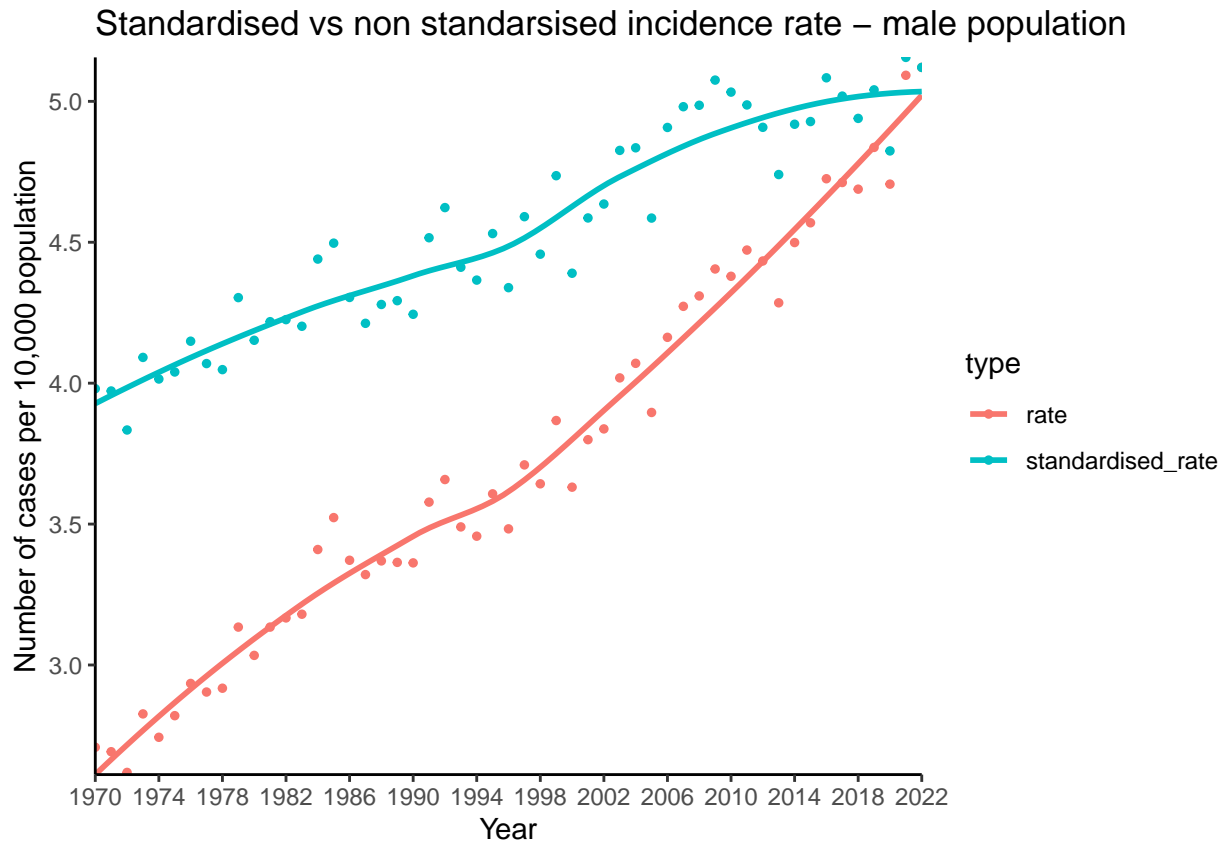
```
cases_pop_ys %>%
  filter(sex == "Female") %>%
  left_join(age_standardised_rates) %>%
  select(year, sex, rate, standardised_rate) %>%
  pivot_longer(cols = c(rate, standardised_rate), names_to = "type") %>%
  ggplot(aes(x = year, y = value, color = type)) +
  # smooth the curve
  geom_smooth(se = FALSE) +
  # non smoothed data
  geom_point(size = 1) +
  theme_classic() +
  coord_cartesian(expand = FALSE) +
  scale_x_continuous(lim = c(1970, 2022), breaks = c(seq(1970, 2022, by = 4))) +
  # add labels and title
  ggtitle("Standardised vs non standardised incidence rate - female population") +
  xlab("Year") + ylab("Number of cases per 10,000 population")
```

Standardised vs non standardised incidence rate – female population



Males

```
cases_pop_ys %>%
  filter(sex == "Male") %>%
  left_join(age_standardised_rates) %>%
  select(year, sex, rate, standardised_rate) %>%
  pivot_longer(cols = c(rate, standardised_rate), names_to = "type") %>%
  ggplot(aes(x = year, y = value, color = type)) +
  # smooth the curve
  geom_smooth(se = FALSE) +
  # non smoothed data
  geom_point(size = 1) +
  theme_classic() +
  coord_cartesian(expand = FALSE) +
  scale_x_continuous(lim = c(1970, 2022), breaks = c(seq(1970, 2022, by = 4))) +
  # add labels and title
  ggtitle("Standardised vs non standardised incidence rate - male population") +
  xlab("Year") + ylab("Number of cases per 10,000 population")
```



Both for females and for males, compared to the original rates, the curve for the standardised rates has a greater value at the intercept and a smaller slope, especially initially. However, as time is increased, the slopes tend to align. The standardized rates are consistently higher than the original rates, with the difference decreasing monotonely over time.

This could be explained by the different structure in the population over the years and the proportional increase in population aged 75 and older, where cancer rates are in general the highest.

Question 12: Direct age standardised rate on predicted data - Poisson model with ages

```
cases_pop_fit <- cases_pop %>%
  select(-rate, -n) %>%
  # replace number of cases with fitted number of cases, which should already be in the right order
  add_column(n_fit = poisson_model_age$fitted.values)

# repeat question 11
age_standardised_rates_fit <- cases_pop_fit %>%
  left_join(standard_pop) %>%
  # add number of cases obtained if population number was ref_pop
  mutate(cases_ref_pop = ref_pop * n_fit/n_pop) %>%
  # for each year, compute total (sum over agegroups) and divide by total reference population
  # keep male and female categories
  group_by(year, sex) %>%
  # sum over agegroups
  summarize(sum_ages = sum(cases_ref_pop, na.rm = TRUE)) %>%
```

```

ungroup() %>%
mutate(standardised_rate_fit = case_when(sex == "Female" ~ sum_ages/tot_female,
                                          sex == "Male" ~ sum_ages/tot_male,
                                          TRUE ~ NA)

      ) %>%
select(-sum_ages) %>%
# rescale
mutate(standardised_rate_fit = 10-4 * standardised_rate_fit)

```

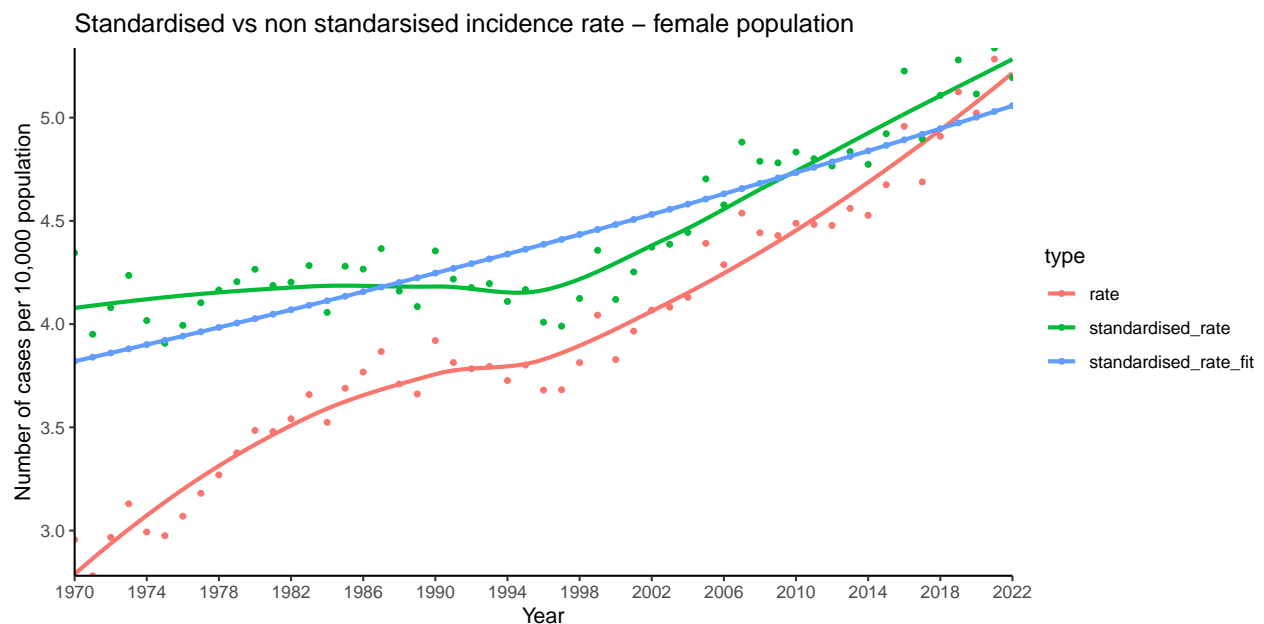
Comparison of standardised rates obtained with fitted data, with standardised rates from the original dataset.

Females

```

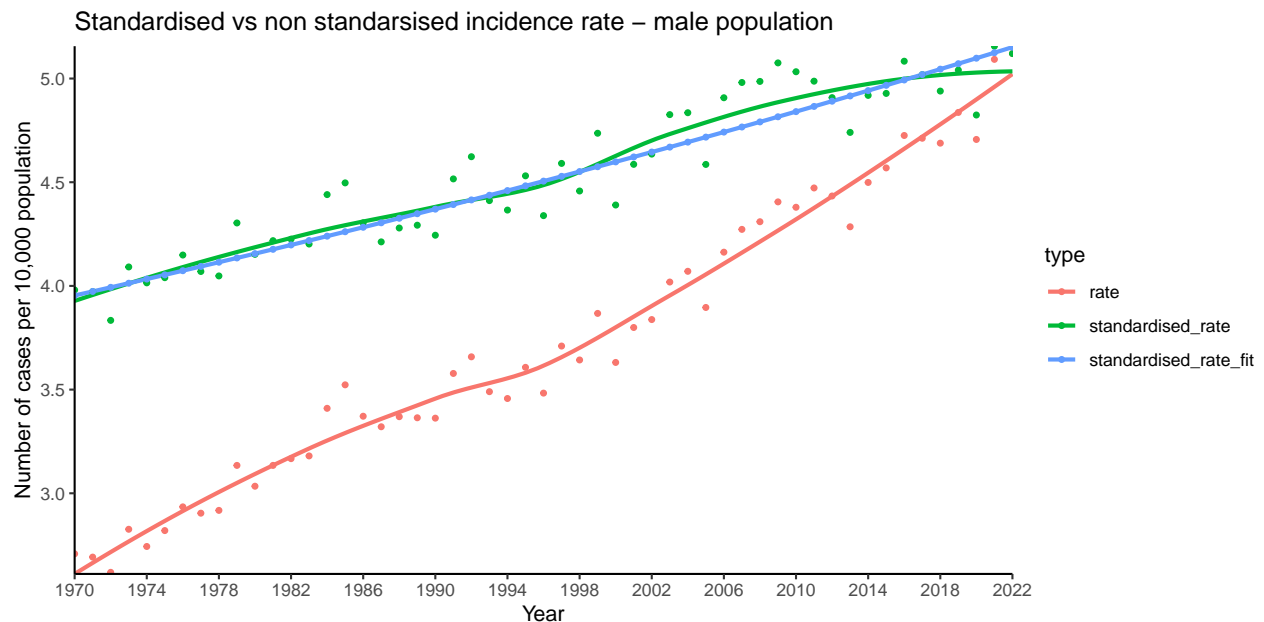
cases_pop_ys %>%
  filter(sex == "Female") %>%
  left_join(age_standardised_rates) %>%
  left_join(age_standardised_rates_fit) %>%
  select(year, sex, rate, standardised_rate, standardised_rate_fit) %>%
  pivot_longer(cols = c(rate, standardised_rate, standardised_rate_fit), names_to = "type") %>%
  #filter(type == "standardised_rate_fit") %>%
  ggplot(aes(x = year, y = value, color = type)) +
  # smooth the curve
  geom_smooth(se = FALSE) +
  # non smoothed data
  geom_point(size = 1) +
  theme_classic() +
  coord_cartesian(expand = FALSE) +
  scale_x_continuous(lim = c(1970, 2022), breaks = c(seq(1970, 2022, by = 4))) +
  # add labels and title
  ggtitle("Standardised vs non standardised incidence rate - female population") +
  xlab("Year") + ylab("Number of cases per 10,000 population")

```



Males

```
cases_pop_ys %>%
  filter(sex == "Male") %>%
  left_join(age_standardised_rates) %>%
  left_join(age_standardised_rates_fit) %>%
  select(year, sex, rate, standardised_rate, standardised_rate_fit) %>%
  pivot_longer(cols = c(rate, standardised_rate, standardised_rate_fit), names_to = "type") %>%
  ggplot(aes(x = year, y = value, color = type)) +
  # smooth the curve
  geom_smooth(se = FALSE) +
  # non smoothed data
  geom_point(size = 1) +
  theme_classic() +
  coord_cartesian(expand = FALSE) +
  scale_x_continuous(lim = c(1970, 2022), breaks = c(seq(1970, 2022, by = 4))) +
  # add labels and title
  ggtitle("Standardised vs non standardised incidence rate - male population") +
  xlab("Year") + ylab("Number of cases per 10,000 population")
```



The curve for the standardised fitted rates appears to follow the standardized original rates quite well for men. For females, however, this is not as clear, which might just be a consequence of the less linear curve for females' original rates with respect to males'.

Question 13: Conclusions

Incidence rates of colon cancer vary significantly across age groups, with older age groups generally having higher incidence rates as seen in the plot "Colon Cancer Incidence Rate by Sex and Age Group Over Time" in question 6.

1. For age group 0-4 and 85-89 the incidence rate is more constant over the time period 1970- 2020 for males than for females.
2. For age group 40-44 there was an increase by a factor of ten or so between the years 2000 and 2020 whereas there was no such change for females.

When we take age into account females tend to have lower incidence rate than males for ages 75 and older: this could be related to the fact, that, for these agegroups, the female population is more numerous than men population.

Over time, there is an overall increase in incidence rates for most age groups, particularly for middle-aged and older adults.

Females consistently show slightly higher crude incidence rates compared to males throughout the observed time period.