Assignment I

Group 2

2024-11-22

Introduction

The objective of this report is to analyze the incidence of colon cancer in Sweden over time, focusing on variations across calendar years, age groups, and sex. By integrating population data with cancer case data, the report aims to compute and interpret incidence rates, exploring patterns and trends. Additionally, the study utilizes statistical modeling, including Poisson regression and age standardization, to provide a deeper understanding of the temporal and demographic aspects of colon cancer incidence. The analysis highlights both observed and modeled incidence rates, contributing to insights into public health trends and cancer epidemiology.

Question 1

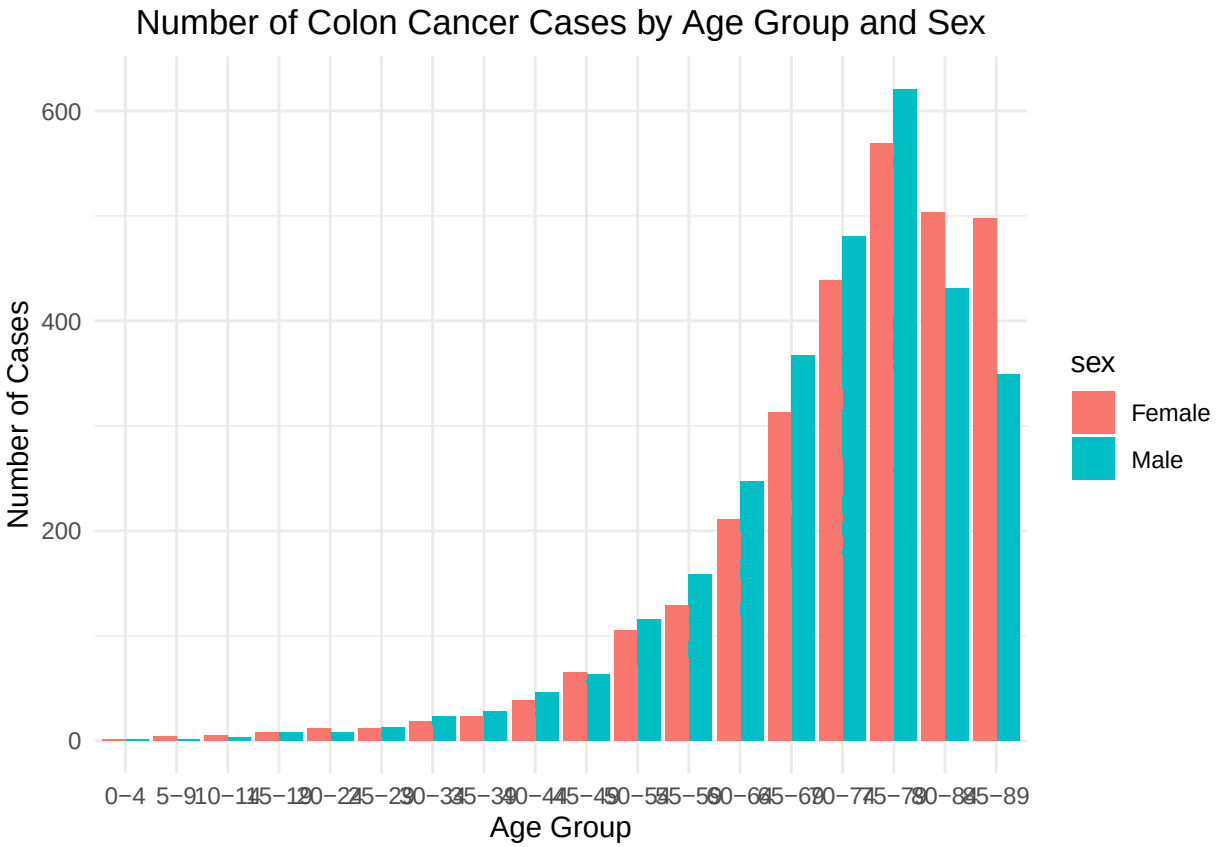
Overview:

Creating a plot showing the number of cases by age group and sex.

Code:

```
path_to_cases <- "cases.tsv"
cases <- read_tsv(path_to_cases, show_col_types = FALSE)

cases$agegroup <- factor(cases$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"))

options(repr.plot.width=40, repr.plot.height=9)
ggplot(cases, aes(x = agegroup, y = n, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Number of Colon Cancer Cases by Age Group and Sex",
       x = "Age Group",
       y = "Number of Cases") +
  theme_minimal()+
  scale_y_continuous(labels = label_comma())+
  theme(plot.title = element_text(hjust = 0.5))
```

Output:



Generated plot visualizing age group distribution.

Analysis:

From the plot we can conclude that the risk of getting a colon cancer increases exponentially above the age approx. 40-44, for both men and women. We can also see a clear trend of the number of cases being slightly higher for men up to the age group 80-84 when the trend reverses.

Question 2

Overview:

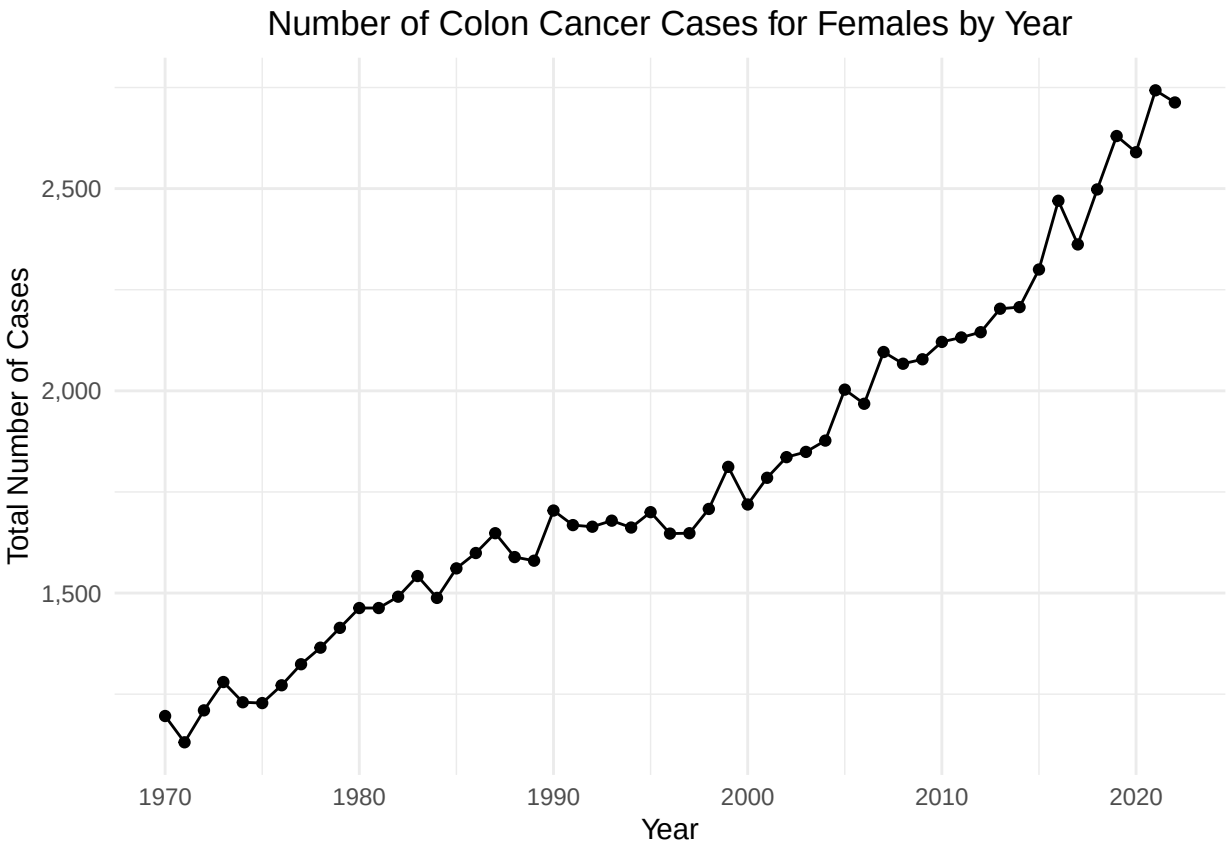
Plotting colon cancer cases across subsequent years.

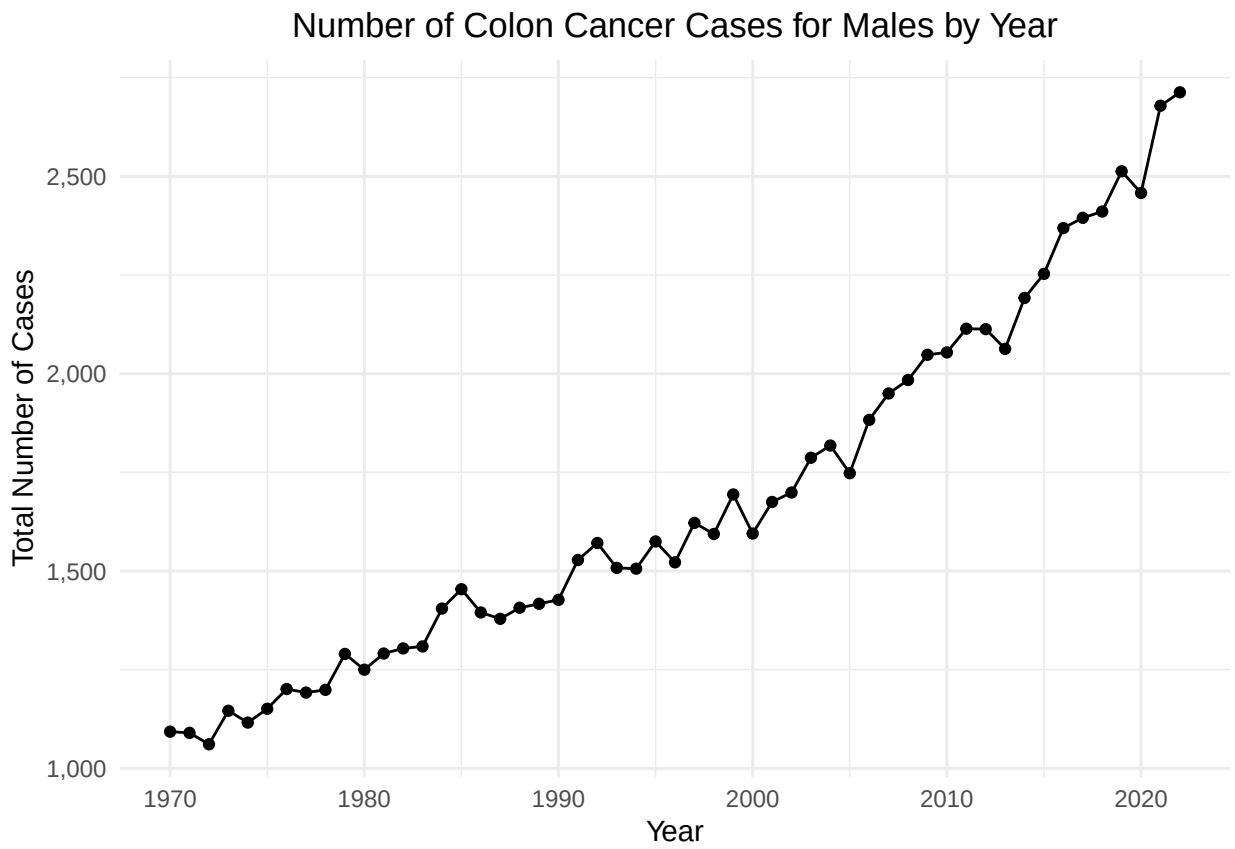
Code:

```r
total_cases_by_year_sex<- cases %>%
  group_by(year, sex) %>%
  summarise(total_cases = sum(n, na.rm = TRUE))

options(repr.plot.width=16, repr.plot.height=9)
ggplot(subset(total_cases_by_year_sex, sex == "Female"), aes(x = year, y = total_cases)) +
  geom_line() +
  geom_point() +
  labs(title = "Number of Colon Cancer Cases for Females by Year",
       x = "Year",
       y = "Total Number of Cases") +
  theme_minimal()+
  scale_y_continuous(labels = label_comma())+
  theme(plot.title = element_text(hjust = 0.5))

options(repr.plot.width=16, repr.plot.height=9)
ggplot(subset(total_cases_by_year_sex, sex == "Male"), aes(x = year, y = total_cases)) +
  geom_line() +
  geom_point() +
  labs(title = "Number of Colon Cancer Cases for Males by Year",
       x = "Year",
       y = "Total Number of Cases") +
  theme_minimal()+
  scale_y_continuous(labels = label_comma())+
  theme(plot.title = element_text(hjust = 0.5))
```

Output:

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


Number of Colon Cancer Cases for Females by Year
Total Number of Cases
2,500
2,000
1,500
1970
1980
1990
2000
2010
2020
Year



Two plots showing the number of cases in each calendar year.

Analysis:

From the two graphs we can see that there exists a trend of increasing colon cancer cases in subsequent calendar years.

Question 3

Overview:

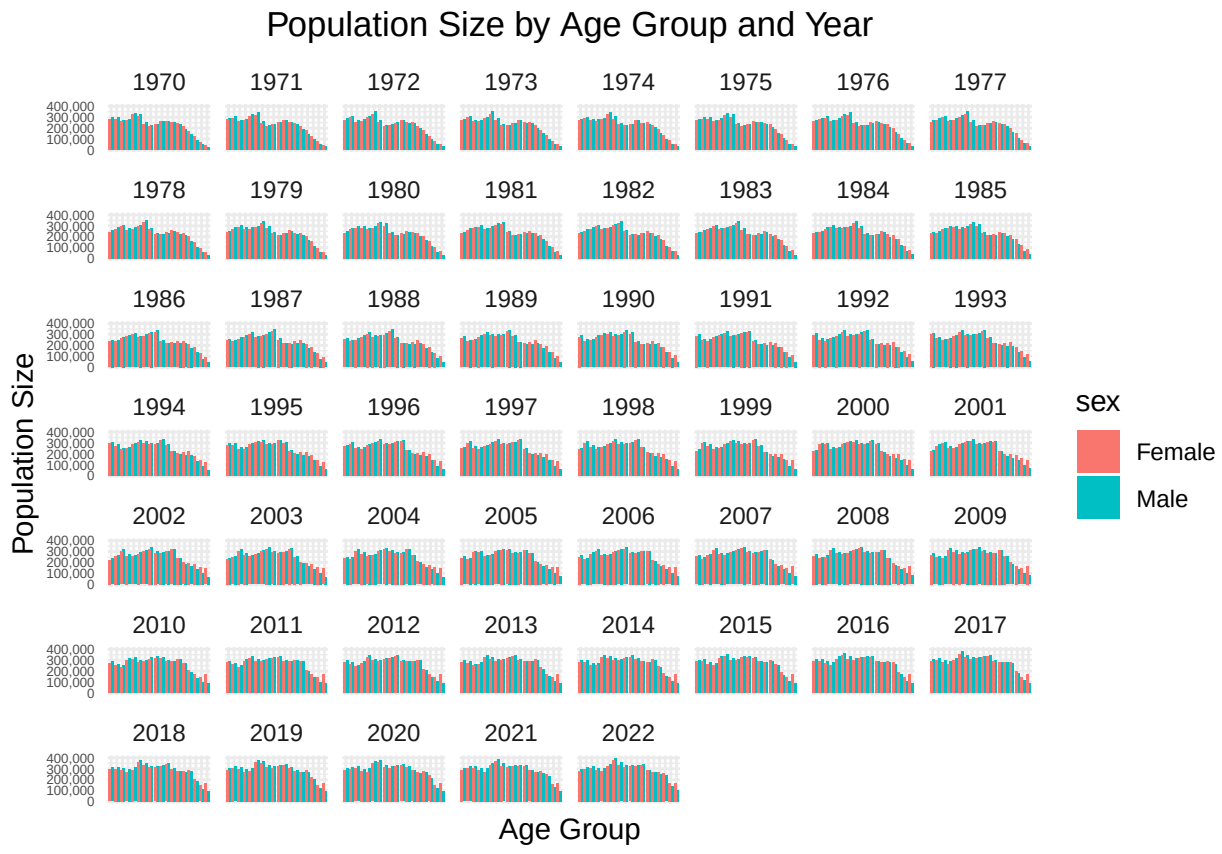
Exploring population data and plotting population size by age group and sex across subsequent years.

Code :

```
path_to_population <- "population.tsv"
population<- read_tsv(path_to_population, show_col_types = FALSE)

population$agegroup <- factor(population$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"))

options(repr.plot.width=32, repr.plot.height=18)
ggplot(population, aes(x = agegroup, y = n_pop, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ year) +
  labs(title = "Population Size by Age Group and Year",
       x = "Age Group",
       y = "Population Size") +
  theme_minimal() +
  scale_y_continuous(labels = label_comma())+
  theme(axis.text.x = element_blank())+
  theme(plot.title = element_text(hjust = 0.5))
```

Output:



A figure showing a plots for every year in the data set. Each of the plots shows population size across different age groups, the bars colors correspond to sex (orange - female, green - male).

Analysis:

Population file includes the same variables and their types as cases file. The age groups, and calendar years are also the same. The difference is in the order of columns, and the year column is reversed in population, relative to cases.

Question 4

Overview:

Merging the data and creating a data frame that shows total number of cases and the total population of males and females in each year.

Code:

```
merged_data <- left_join(cases, population, by = c("agegroup", "year", "sex"))

summary_data <- merged_data %>%
  group_by(year, sex) %>%
  summarise(
    total_cases = sum(n, na.rm = TRUE),
    total_population = sum(n_pop, na.rm = TRUE)
  )
```

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

```
head(summary_data)
```

```
## # A tibble: 6 x 4
## # Groups:   year [3]
##    year sex    total_cases total_population
##   <dbl> <chr>        <dbl>            <dbl>
## 1  1970 Female        1196          4045318
## 2  1970 Male          1093          4035911
## 3  1971 Female        1131          4066592
## 4  1971 Male          1090          4048573
## 5  1972 Female        1210          4077814
## 6  1972 Male          1061          4051315
```

Output:

New data frame that stores information merged from both of the analyzed data sets.

Question 5

Overview:

Creating a new data frame that will show incidence rate among both sexes in each year.

Code:

```
merged_data <- merged_data %>%
  mutate(incidence_rate = n/ n_pop)

head(merged_data)
```

```
## # A tibble: 6 x 6
##   agegroup  year sex       n  n_pop incidence_rate
##   <fct>    <dbl> <chr> <dbl>  <dbl>          <dbl>
## 1 0-4       2022 Male      0 296183     0
## 2 5-9       2022 Male      0 319820     0
## 3 10-14     2022 Male      1 325003     0.00000308
## 4 15-19     2022 Male      8 310539     0.0000258
## 5 20-24     2022 Male      5 310354     0.0000161
## 6 25-29     2022 Male      4 342974     0.0000117
```

```
summary_data <- summary_data %>%
  mutate(incidence_rate = total_cases / total_population)

head(summary_data)
```

```
## # A tibble: 6 x 5
## # Groups:   year [3]
##    year sex    total_cases total_population incidence_rate
##   <dbl> <chr>        <dbl>            <dbl>          <dbl>
## 1  1970 Female        1196          4045318       0.000296
## 2  1970 Male          1093          4035911       0.000271
## 3  1971 Female        1131          4066592       0.000278
## 4  1971 Male          1090          4048573       0.000269
## 5  1972 Female        1210          4077814       0.000297
## 6  1972 Male          1061          4051315       0.000262
```

Output:

A dataframe with information on the number of cases, population and incidence year for men and women in each year.

Analysis:

Incidence rate is the number of new cases of the outcome divided by the total person-time at risk, for a specific follow-up period. Here we simply divide the number of cases by the population size, without

accounting for the time. For this type of data where we do not know the person-time at risk, it appears to be an appropriate way of calculating an incidence rate. However, this way should provide a reasonable estimate when, for example, the population is relatively stable over the time period (we can see from the plots in Question 3 that the population changes over the given time - we have more people in the elderly age group as the years increase). This suggests that this data is not suitable to infer the incidence rate.

Question 6

Overview:

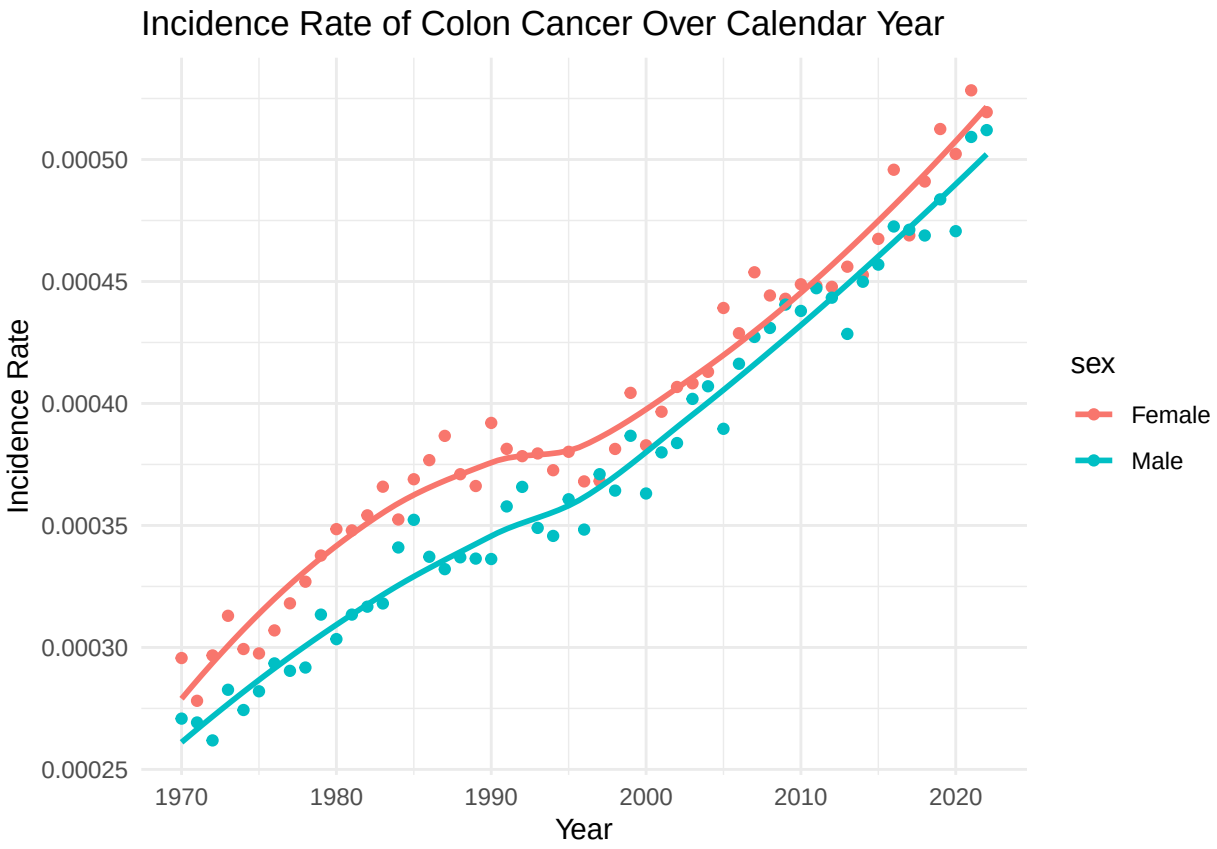
Plotting the incidence rate of colon cancer over calendar time and apply a smoother, separately by males and females. Creating a graph of incidence rates over calendar year by sex and age group, and apply a smoother.

Code:

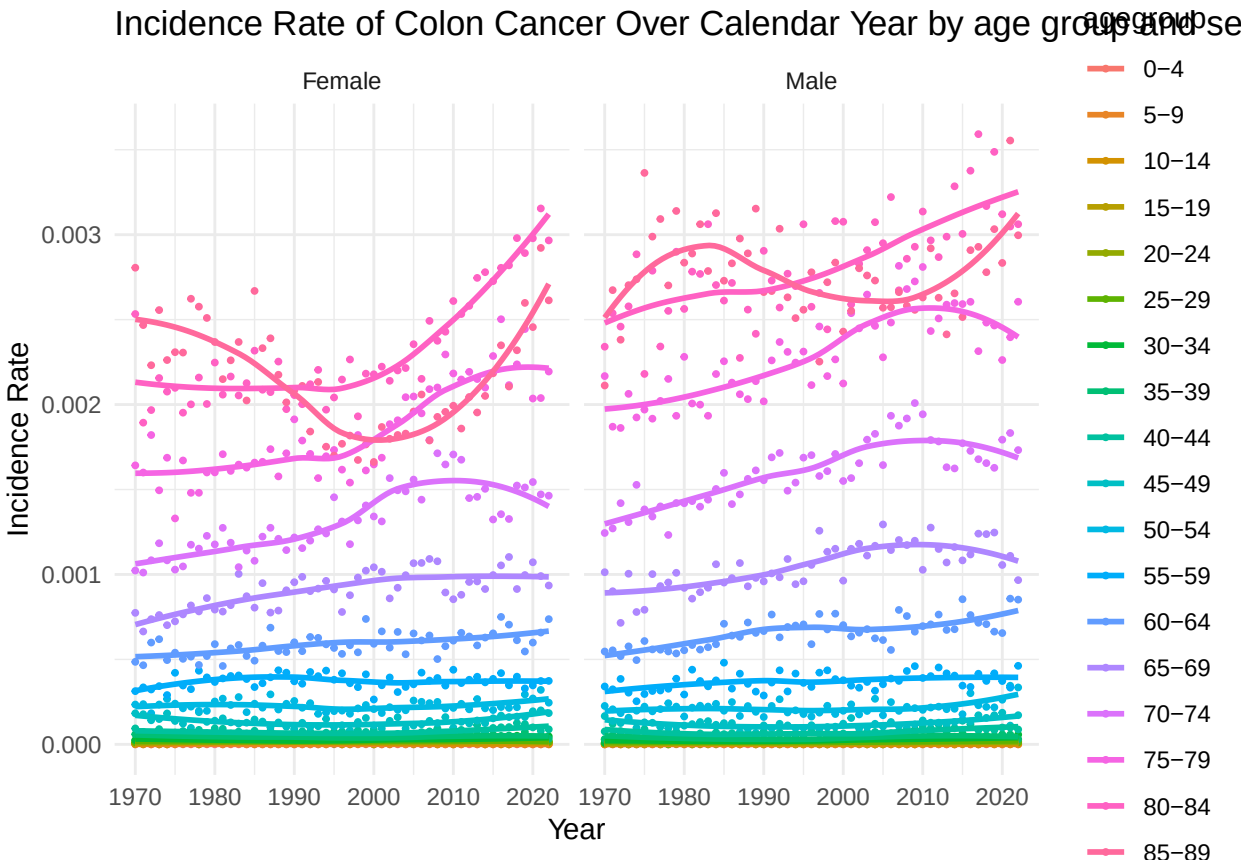
```
ggplot(summary_data, aes(x = year, y = incidence_rate, color = sex)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Incidence Rate of Colon Cancer Over Calendar Year",
       x = "Year",
       y = "Incidence Rate") +
  theme_minimal()

ggplot(merged_data, aes(x = year, y = incidence_rate, color = agegroup)) +
  geom_point(cex = 0.7) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(title = "Incidence Rate of Colon Cancer Over Calendar Year by age group and sex",
       x = "Year",
       y = "Incidence Rate") +
  theme_minimal() +
  facet_wrap(~ sex)
```

Output:

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




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



A graph showing the smoothed incidence rate trends for males and females over time. Faceted graphs showing incidence rate trends by age group and sex over calendar years.

Analysis:

Overall Trends:

The smoothed curves reveal increasing incidence rates for both sexes over the observed time frame. Differences in incidence rates between males and females are consistent over time, with females generally having higher rates.

Age Group-Specific Trends:

There is a noticeable rise in incidence rates over time in most age groups, particularly in the older populations. The youngest age groups (e.g., 0-4, 5-9, 10-14) maintain consistently low incidence rates throughout the years, with little variation. For males, the incidence rates in the highest age groups (e.g., 70-74 and above) are more pronounced compared to females. The higher the age group, the higher the probability of colon cancer, but for age group 85-89, the incidence of colon cancer fluctuated with the increase of years, and after around 1990, the incidence of colon cancer in this age group was lower than that in age group 80-84 The trends for middle-aged groups (e.g., 40-44, 50-54) exhibit moderate increases over the years, with smoother and less steep curves compared to older groups. This suggests a gradual risk increase with age but highlights that the exponential increase occurs predominantly in the elderly.

Question 7

Overview:

Fitting a suitable Poisson model with the total number of cases as dependent variable, using the population size as an offset, and calendar year and sex as independent variables.

Code:

```
poisson_model <- glm(total_cases ~ year + sex + offset(log(total_population)),
                     data = summary_data,
                     family = poisson)

summary(poisson_model)
```

```
##
## Call:
## glm(formula = total_cases ~ year + sex + offset(log(total_population)),
##     family = poisson, data = summary_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
```

Output:

A detailed summary of the Poisson regression model, showing the coefficients, standard errors, z-values, and significance levels for each predictor.

Analysis:

Yearly Trend: The positive coefficient for year suggests a consistent increase in colon cancer incidence rates over time. Sex-Specific Differences: While males generally have higher crude incidence rates than females, the negative coefficient for sex Male suggests that, after controlling for year and population size, the adjusted incidence rate is slightly lower for males.

Question 8

Overview:

Based on the poisson regression model, calculate the incidence rates in 1970 and 2020 for males and females, and discuss the assumptions about how rates change over time and between sexes.

Code:

```
predict_data <- data.frame(year = c(1970, 2020),
                           sex = rep(c("Male", "Female"), each = 2),
                           total_population = 100000)

predict_data$predicted_cases <- predict(poisson_model,
                                        newdata = predict_data,
                                        type = "response")

predict_data$incidence_rate <- (predict_data$predicted_cases / predict_data$total_population)
print(predict_data)
```

```
##   year    sex total_population predicted_cases incidence_rate
## 1 1970   Male            1e+05        27.73751   0.0002773751
## 2 2020   Male            1e+05        47.94199   0.0004794199
## 3 1970 Female            1e+05        29.33271   0.0002933271
## 4 2020 Female            1e+05        50.69918   0.0005069918
```

Output:

A data frame with year, sex, and total population in the first three columns and the predicted number of cases as well as the corresponding incidence rate in the last two columns.

Analysis:

From the result, the incidence rate in 1970 among males and females was 0.0002773751 and 0.0002933271, separately. Moreover, the incidence rate in 2020 among males and females was 0.0004794199 and 0.0004794199, separately. The assumption for year is that the incidence rate changes linearly over the calendar year, and the assumption for the sex is that there is no interaction between sex and year, which means that the ratio of the incidence rate of male and female is an unchanged constant.

Question 9

Overview:

Fit a Poisson regression model that adjusts for age groups and calculate the incidence rates for males and females in the 70-74 age group for 1970 and 2020.

Code:

```
poisson_model_age <- glm(n ~ year + sex + agegroup + offset(log(n_pop)),
                         data = merged_data,
                         family = poisson)

summary(poisson_model_age)
```

```
##
## Call:
## glm(formula = n ~ year + sex + agegroup + offset(log(n_pop)),
##     family = poisson, data = merged_data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -27.338166   0.771867 -35.418  < 2e-16 ***
## year            0.005400   0.000155  34.851  < 2e-16 ***
## sexMale         0.155476   0.004687  33.172  < 2e-16 ***
## agegroup5-9     2.841194   0.727020   3.908 9.31e-05 ***
## agegroup10-14   4.304576   0.711734   6.048 1.47e-09 ***
## agegroup15-19   5.058536   0.709256   7.132 9.88e-13 ***
## agegroup20-24   5.263833   0.708782   7.427 1.11e-13 ***
## agegroup25-29   5.452471   0.708415   7.697 1.40e-14 ***
## agegroup30-34   5.912646   0.707926   8.352  < 2e-16 ***
## agegroup35-39   6.397141   0.707621   9.040  < 2e-16 ***
## agegroup40-44   6.962570   0.707402   9.842  < 2e-16 ***
## agegroup45-49   7.506562   0.707277  10.613  < 2e-16 ***
## agegroup50-54   8.052154   0.707205  11.386  < 2e-16 ***
## agegroup55-59   8.568783   0.707164  12.117  < 2e-16 ***
## agegroup60-64   9.092283   0.707139  12.858  < 2e-16 ***
## agegroup65-69   9.541911   0.707126  13.494  < 2e-16 ***
## agegroup70-74   9.944553   0.707120  14.063  < 2e-16 ***
## agegroup75-79  10.287697   0.707117  14.549  < 2e-16 ***
## agegroup80-84  10.504311   0.707121  14.855  < 2e-16 ***
## agegroup85-89  10.407244   0.707130  14.718  < 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: 406971  on 1907  degrees of freedom
## Residual deviance:   3472  on 1888  degrees of freedom
## AIC: 12620
##
## Number of Fisher Scoring iterations: 6
```

```
predict_data_age <- data.frame(year = c(1970, 2020),
                               agegroup = '70-74',
                               sex = rep(c("Male", "Female"), each = 2),
                               n_pop = 100000)

predict_data_age$predicted_cases <- predict(poisson_model_age,
```

```
                                        newdata = predict_data_age,
                                        type = "response")

predict_data_age$incidence_rate <- (predict_data_age$predicted_cases / predict_data_age$n_pop)
print(predict_data_age)
```

```
##   year agegroup    sex n_pop predicted_cases incidence_rate
## 1 1970    70-74   Male 1e+05        136.0602    0.001360602
## 2 2020    70-74   Male 1e+05        178.2354    0.001782354
## 3 1970    70-74 Female 1e+05        116.4685    0.001164685
## 4 2020    70-74 Female 1e+05        152.5708    0.001525708
```

Output:

A detailed summary information of the new poisson model. A data frame with year, age group, sex, and population in the first four columns and the predicted number of cases as well as the corresponding incidence rate in the last two columns.

Analysis:

From the result, the incidence rate in 1970 in age group 70-74 among males and females was 0.001360602 and 0.001164685, separately. Moreover, the incidence rate in 2020 in age group 70-74 among males and females was 0.001782354 and 0.001525708, separately.

Question 10

Overview:

Refit the above model using splines for the effect of calendar year and age group. Plotting graphs to show the incidence rate across calendar time for males and females at ages 52, 72 and 87 and compare it with the known observed values for age group 50-54, 70-74 and 85-89.

Code:

1. Refit the model using splines:

- Remove the assumption that the pattern across calendar year is the same across age and sex by adding relation between calendar year, age and sex using natural splines.
- Natural Splines were used here since the relation between year, age and sex are not highly complex and unpredictable, and also for a smoother pattern and a cleaner fit.
- The degrees of freedom were chosen to be 4 as it had the least AIC compared to the models with other degrees of freedom and it did not pose a risk of over-fitting by requiring higher parameters for a smoother curve.
- A three-way interaction between year, sex and age-mid was included in the poisson model to capture more complexity and nuanced interactions between year, sex, and age and the subtle patterns in the data.

```r
# Convert agegroup to midpoints
age_midpoints <- c(2.5, 7.5, 12.5, 17.5, 22.5, 27.5, 32.5, 37.5, 42.5, 47.5,
                   52.5, 57.5, 62.5, 67.5, 72.5, 77.5, 82.5, 87.5)
names(age_midpoints) <- levels(cases$agegroup)
merged_data <- merged_data %>%
  mutate(age_mid = age_midpoints[as.character(agegroup)])

# Fit Poisson model with splines
spline_model <- glm(n ~ ns(year, df = 4) * sex * ns(age_mid, df = 4),
                    data = merged_data,
                    offset = log(n_pop),
                    family = poisson)

print(spline_model$aic)
```

```
## [1] 12693.97
```

Please refer to the Appendix for the entire summary for the spline model.

2. Plot the graphs with modeled vs observed incidence rate across calendar time:

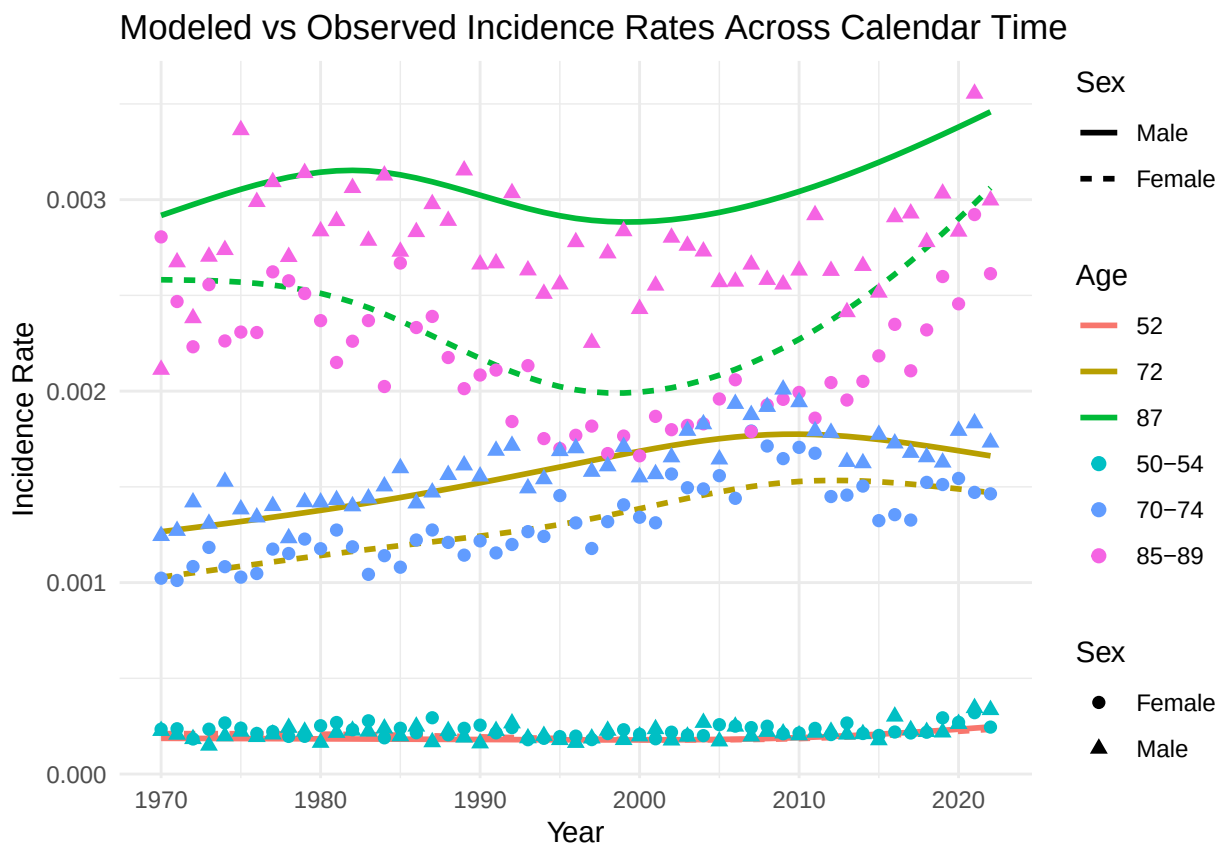
- First get the data for prediction from the newly fitted spline model and extract the observed data for the age groups 50-54, 70-74, 85-89.

```r
# Data for prediction
predict_data <- expand.grid(
  year = seq(min(merged_data$year), max(merged_data$year), by = 1),
  age_mid = c(52, 72, 87),  # Specific ages for prediction
  sex = c("Male", "Female"),
  n_pop = 100000  # Assume a constant population for prediction
)

predict_data$predicted_cases <- predict(spline_model, newdata = predict_data,
                                        type = "response")
predict_data$incidence_rate <- predict_data$predicted_cases / predict_data$n_pop

# Extract observed data for age groups 50-54, 70-74, 85-89
observed_data <- merged_data %>%
  filter(agegroup %in% c("50-54", "70-74", "85-89")) %>%
  group_by(year, sex, agegroup) %>%
  summarise(incidence_rate = sum(n) / sum(n_pop), .groups = 'drop')
```

Output:

```r
# Plot modeled vs. observed incidence rates
ggplot() +
  geom_line(data = predict_data, aes(x = year, y = incidence_rate,
                                     color = factor(age_mid), linetype = sex),
            linewidth = 1) +
  geom_point(data = observed_data, aes(x = year, y = incidence_rate,
```

```
                                    shape = sex, color = factor(agegroup)),
            size = 2) +
labs(
  title = "Modeled vs Observed Incidence Rates Across Calendar Time",
  x = "Year",
  y = "Incidence Rate",
  color = "Age",
  linetype = "Sex",
  shape = "Sex"
) +
theme_minimal()
```



Analysis:

1. Overall Trends: - The modeled incidence rates (lines) show a gradual increase across calendar years for both sexes and age groups retaining the trend of drop in the incidence rate in females of age 85-89 between the year 2000 and 2010. - The observed values (points) generally align with the modeled rates but exhibit more variability, especially in older age groups.

2. Differences by Age: - Incidence rates are consistently higher for older age groups (e.g., 87) compared to younger ones (e.g., 52), as clearly shown by both the modeled and observed rates. - The gap between modeled and observed rates is slightly larger for older age groups, suggesting potential model misfit or variability in the data.

3. Sex Differences: - For all age groups, the incidence rates are higher for males (solid lines, triangles)

compared to females (dashed lines, circles), except for age 52, which are almost entirely aligned for male and female.

4. Model Fit for Age Groups: - The modeled rates for ages 52 and 72 fit the observed data closely, indicating a good representation of trends, except for age 87, which generally captured the increasing trend in the modeled rates, had some noticeable deviations from the observed points for females.

Question 11

Overview:

The task is to calculate direct age-standardized incidence rates for colon cancer across calendar years and sexes, using the population age distribution from 2022 as the reference (standard population). The objective is to create a graph of the age-standardized incidence rates and compare it with the non-age-standardized (crude) rates for males and females over time.

Code:

To solve this, there are a series of steps that involves significant data pre-processing.

1. Define a Standard Population:

Use the age distribution from 2022 as the reference.

```
# Summarize the age distribution in 2022
standard_population <- population %>%
  filter(year == 2022) %>%
  group_by(sex, agegroup) %>%
  summarise(Population_2022 = sum(n_pop), .groups = "drop")

head(standard_population)
```

```
## # A tibble: 6 x 3
##   sex    agegroup Population_2022
##   <chr>  <fct>              <dbl>
## 1 Female 0-4               280184
## 2 Female 5-9               301335
## 3 Female 10-14             306568
## 4 Female 15-19             292308
## 5 Female 20-24             275136
## 6 Female 25-29             323678
```

2. Calculate Crude Rates:

Determine the number of cases per unit population for each year, age group, and sex.

```
# Merge cases and population data
merged_data <- cases %>%
  left_join(population, by = c("year", "agegroup", "sex"))
```

```r
# Calculate crude incidence rates
merged_data <- merged_data %>%
  mutate(Crude_Incidence_Rate = n / n_pop)

head(merged_data)
```

```
## # A tibble: 6 x 6
##   agegroup  year sex       n  n_pop Crude_Incidence_Rate
##   <fct>    <dbl> <chr> <dbl>  <dbl>                <dbl>
## 1 0-4       2022 Male      0 296183           0
## 2 5-9       2022 Male      0 319820           0
## 3 10-14     2022 Male      1 325003           0.00000308
## 4 15-19     2022 Male      8 310539           0.0000258
## 5 20-24     2022 Male      5 310354           0.0000161
## 6 25-29     2022 Male      4 342974           0.0000117
```

3. Standardize Rates:

Weight the crude rates by the 2022 age distribution to compute age-standardized rates.

```r
standardized_rates <- merged_data %>%
  filter(!is.na(Crude_Incidence_Rate)) %>%
  left_join(standard_population, by = c("sex", "agegroup")) %>%
  group_by(year, sex) %>%
  summarise(
    Standardized_Rate = sum(Crude_Incidence_Rate * Population_2022) / sum(Population_2022),
    .groups = "drop"
  )

head(standardized_rates)
```

```
## # A tibble: 6 x 3
##    year sex    Standardized_Rate
##   <dbl> <chr>              <dbl>
## 1  1970 Female          0.000435
## 2  1970 Male            0.000398
## 3  1971 Female          0.000395
## 4  1971 Male            0.000397
## 5  1972 Female          0.000408
## 6  1972 Male            0.000383
```

4. Visualization:

Compare age-standardized and crude incidence rates across calendar years to evaluate the impact of standardization and reveal true trends.

We first compute the crude incidence rates and group by year and sex, similar to how we computed 'standardized_rates'.

```r
crude_rates <- merged_data %>%
  group_by(year, sex) %>%
  summarise(
    Crude_Rate = sum(n) / sum(n_pop),
    .groups = "drop"
  )
```

In order to compare the age-standardized rates vs the non-standardized rates, we combine the tables 'standardized_rates' and 'crude_rates' and create distinct columns to distinguish between each other.

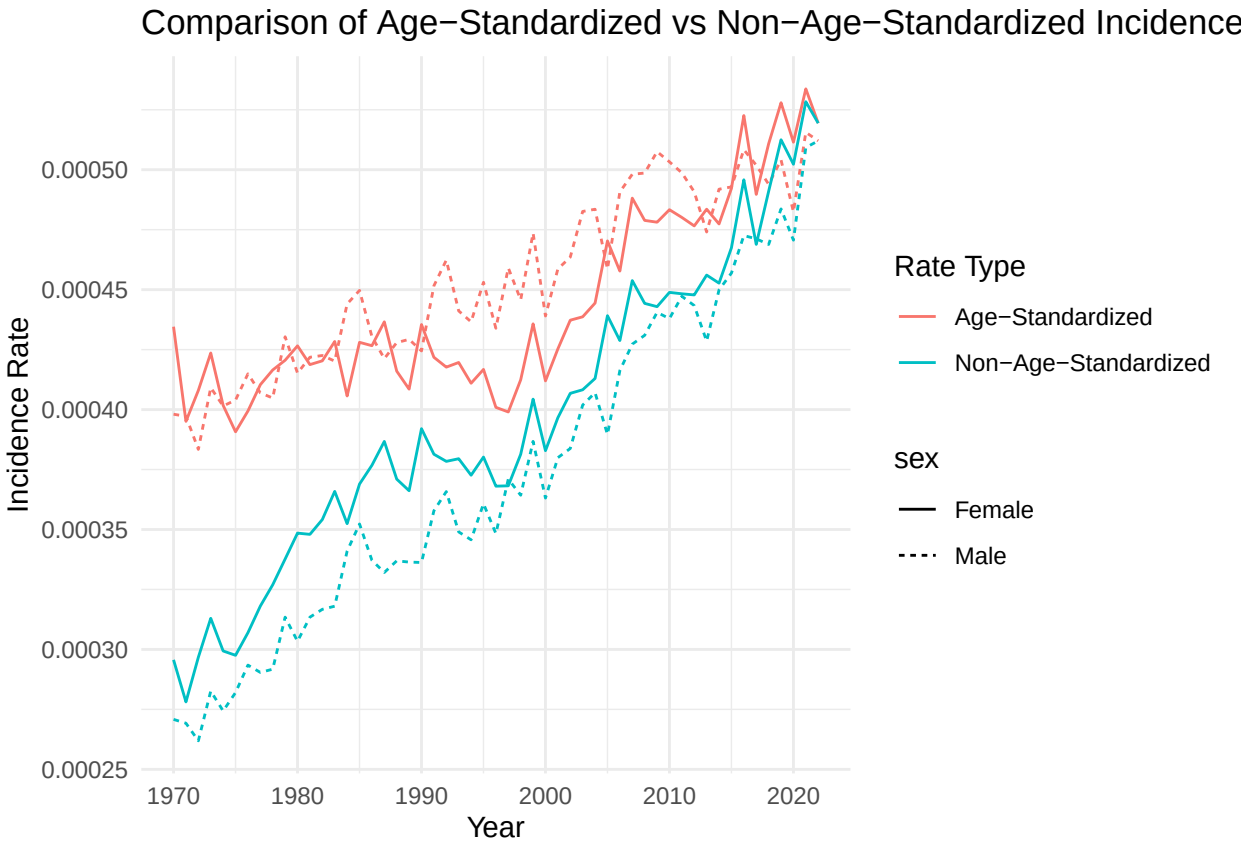
```
comparison_data <- crude_rates %>%
  rename(Rate = Crude_Rate) %>%
  mutate(Type = "Non-Age-Standardized") %>%
  bind_rows(
    standardized_rates %>%
      rename(Rate = Standardized_Rate) %>%
      mutate(Type = "Age-Standardized")
  )

head(comparison_data)
```

```
## # A tibble: 6 x 4
##    year sex        Rate Type
##   <dbl> <chr>     <dbl> <chr>
## 1  1970 Female 0.000296 Non-Age-Standardized
## 2  1970 Male   0.000271 Non-Age-Standardized
## 3  1971 Female 0.000278 Non-Age-Standardized
## 4  1971 Male   0.000269 Non-Age-Standardized
## 5  1972 Female 0.000297 Non-Age-Standardized
## 6  1972 Male   0.000262 Non-Age-Standardized
```

Finally, we plot the graph using this newly created table.

```
ggplot(comparison_data, aes(x = year, y = Rate, color = Type, linetype = sex)) +
  geom_line() +
  labs(
    title = "Comparison of Age-Standardized vs Non-Age-Standardized Incidence Rates",
    x = "Year",
    y = "Incidence Rate",
    color = "Rate Type"
  ) +
  theme_minimal()
```

Output:



Analysis:

1. Difference Between Standardized and Crude Rates:

- The age-standardized rates are generally higher than the crude rates for both sexes, particularly in recent years.
- This suggests that the population's age structure has shifted, with an increasing proportion of older individuals (who have a higher incidence of colon cancer).

2. Trend Over Time:

- Both standardized and crude rates show an increasing trend over time, indicating a rising incidence of colon cancer from 1970 to 2020.
- The steeper increase in crude rates reflects the influence of population aging.

3. Sex-Specific Differences:

- Males consistently exhibit slightly higher incidence rates than females across all years in both age-standardized and crude measures.
- This could be attributed to differences in biological, behavioral, or exposure-related factors.

Conclusion:

Age-standardization adjusts for changes in the population's age distribution, providing a clearer picture of temporal trends. The increasing standardized rates suggest a true rise in colon cancer incidence beyond the effects of aging.

Question 12

Overview:

The aim is to estimate age-standardized rates based on predicted rates from the regression model. This involves using the predicted values from a model (e.g., Poisson regression) and applying the age distribution in 2022 for standardization.

Code:

1. Data Pre-Processing

We first perform the tables from our source tables for prediction. We combine the cases data set (number of cancer cases) with the population data set (population at risk) by matching year, agegroup, and sex. This table is then joined with the standard_population table from Q11.

```
merged_data12 <- cases %>%
  left_join(population, by = c("year", "agegroup", "sex"))

merged_data12 <- merged_data12 %>%
  left_join(standard_population, by = c("sex", "agegroup"))
```

2. Poisson Model

Here, a Poisson regression model is fitted to predict the number of cases (n) based on year, sex, and agegroup.

```
poisson_model <- glm(
  n ~ year + sex + agegroup,
  data = merged_data12,
  family = poisson
)
```

Using this newly fitted model, we predict the number of cases for each row in merged_data12. However, do note that this is not the rate.

```
merged_data12 <- merged_data12 %>%
  mutate(
    Predicted_Count = predict(poisson_model, newdata = ., type = "response")
  )
```

3. Compute Age-standardized Rates

The Predicted rates are standardized using the 2022 standard population (Population_2022).

$$\text{Age-Standardized Rate} = \frac{\sum(\text{Predicted Rate} \times \text{Standard Population})}{\sum(\text{Standard Population})}$$

We also compute the Direct Standardized rates similar to the previous question.

```
standardized_rates_predicted <- merged_data12 %>%
  group_by(year, sex) %>%
  summarise(
    Age_Standardized_Rate = sum(Predicted_Count / n_pop * Population_2022) / sum(Population_2022),
    .groups = "drop"
  )

direct_standardized_rates <- merged_data12 %>%
  group_by(year, sex) %>%
  summarise(
    Direct_Standardized_Rate = sum(n / n_pop * Population_2022) / sum(Population_2022),
    .groups = "drop"
  )
```

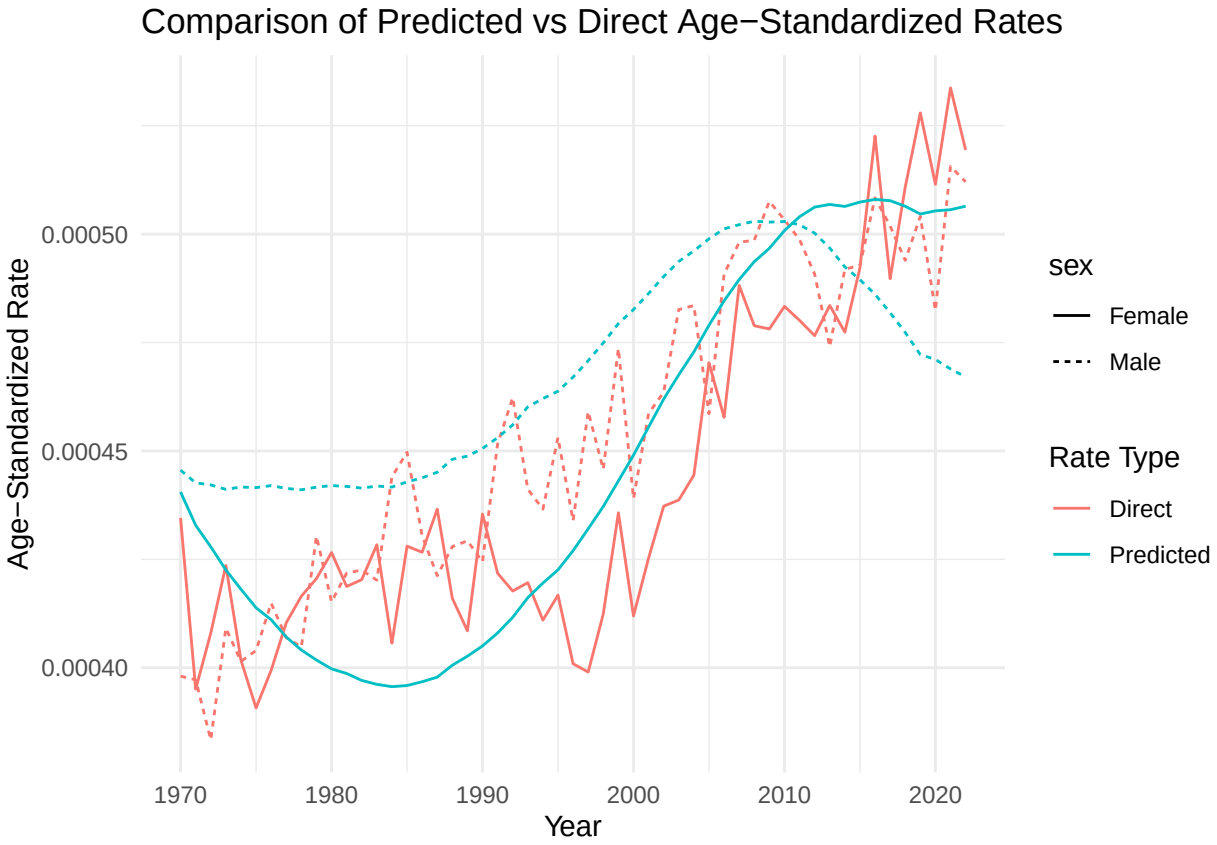
4. Comparison of Predicted vs Direct Age-Standardized Rates

We create a single table by combining the above two tables. In order to eliminate data redundancy, we use bind_rows instead of joining them together. This way, we ensure the rates remain independent given the predictor variables. We introduce a new column to distinguish between the Predicted and Direct rates.

Using this table, we plot the two different types of rates against the year to visualize the trend over time.

```
comparison_data <- standardized_rates_predicted %>%
  rename(Rate = Age_Standardized_Rate) %>%
  mutate(Type = "Predicted") %>%
  bind_rows(
    direct_standardized_rates %>%
      rename(Rate = Direct_Standardized_Rate) %>%
      mutate(Type = "Direct")
  )

ggplot(comparison_data, aes(x = year, y = Rate, color = Type, linetype = sex)) +
  geom_line() +
  labs(
    title = "Comparison of Predicted vs Direct Age-Standardized Rates",
    x = "Year",
    y = "Age-Standardized Rate",
    color = "Rate Type"
  ) +
  theme_minimal()
```

Output:



Analysis:

This graph compares direct age-standardized rates (calculated using observed data) and predicted age-standardized rates (based on a Poisson regression model) across calendar years (1970–2020) for males and females.

1. Alignment Between Predicted and Direct Rates:

- The predicted rates (cyan lines) closely follow the overall trend of the direct rates (red lines), indicating the model's ability to capture the underlying patterns in colon cancer incidence.
- While direct rates exhibit more year-to-year variability, the predicted rates provide a smoother representation of the trend.

2. Sex-Specific Differences:

- Males (dotted lines) consistently have higher incidence rates compared to females (solid lines) throughout the study period. This pattern is evident in both predicted and direct rates.
- The gap between male and female incidence rates remains relatively stable over time.

3. Trend Over Time:

- Both predicted and direct rates suggest an initial decline in colon cancer incidence from 1970 to the early 1980s.

- After this decline, rates show a steady increase from the 1980s onward, continuing through 2020. This trend is observed for both males and females.

4. Fluctuations in Direct Rates:

- The direct rates show noticeable year-to-year fluctuations, particularly in earlier years (e.g., 1970–1980). This variability may reflect sampling noise or demographic changes in smaller population subgroups.

5. Long-Term Increase:

- The increase in incidence rates after the 1980s reflects a persistent upward trend, likely influenced by changes in risk factors, diagnostic practices, or population health.

Question 13

Conclusion regarding the pattern of colon cancer incidence across calendar years:

1. Long-Term Increase in Incidence Rates

- Both observed and modeled data show a persistent increase in colon cancer incidence rates after the 1980s, likely driven by dietary changes, sedentary lifestyles, and advancements in diagnostic techniques. This trend highlights the growing burden of the disease over time.

2. Age-Specific Risks

- Incidence rates increase exponentially with age, with the highest rates observed in the 85-89 age group. This pattern reflects the cumulative nature of cancer risk and the role of aging as a primary risk factor.

3. Sex Differences in Incidence

- Males consistently exhibit higher incidence rates compared to females across all age groups and calendar years. This difference suggests potential biological, hormonal, or behavioral risk factors specific to males, warranting targeted research and prevention.

4. Temporal Changes in Risk for Older Adults

- The fluctuations observed in the 85-89 age group, especially for females, between 2000 and 2010, might indicate changes in diagnostic practices, screening policies, or survival bias among elderly populations.

5. Initial Decline in Rates (1970–1980s)

- A decline in colon cancer incidence rates during the early years (1970–1980) may reflect early public health efforts, dietary shifts, or under-diagnosis due to less advanced screening techniques.

6. Impact of Population Aging

- The increasing crude incidence rates partly result from demographic shifts, with a growing proportion of older individuals in the population. Age-standardized rates reveal that this trend is not solely due to aging, emphasizing the role of other factors like lifestyle and screening.

7. Modeling Effectiveness

- The use of splines to model interactions between calendar year, age, and sex effectively captures nuanced patterns, such as differing trends in males and females and variability across age groups. Predicted rates align closely with observed data, validating the model.

8. Public Health Implications

- The rise in age-standardized incidence rates suggests a true increase in colon cancer risk over time, beyond population aging. This calls for continued investment in preventive measures like promoting healthy diets, physical activity, and reducing smoking and alcohol use.

9. Diagnostic and Screening Trends

- The upward trends in incidence rates for middle-aged groups (e.g., 50-54, 70-74) suggest improved early detection through widespread screening programs. However, this also indicates a growing need for resources to manage detected cases.

10. Need for Targeted Interventions

- The consistently higher incidence rates in males and older adults emphasize the need for targeted interventions focusing on these high-risk groups. Sex-specific prevention strategies and increased access to screening for older populations are critical for reducing the disease burden.

Appendix:

```
summary(spline_model)
```

```
##
## Call:
## glm(formula = n ~ ns(year, df = 4) * sex * ns(age_mid, df = 4),
##     family = poisson, data = merged_data, offset = log(n_pop))
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                -14.034427   0.512991 -27.358
## ns(year, df = 4)1                            1.045770   0.590968   1.770
## ns(year, df = 4)2                           -0.914099   0.658999  -1.387
## ns(year, df = 4)3                            1.667833   1.260395   1.323
## ns(year, df = 4)4                           -0.857349   0.552948  -1.551
```

```
## sexMale                                          -2.609492   1.038435  -2.513
## ns(age_mid, df = 4)1                              4.966499   0.496942   9.994
## ns(age_mid, df = 4)2                              6.128079   0.357984  17.118
## ns(age_mid, df = 4)3                             10.087482   1.049311   9.613
## ns(age_mid, df = 4)4                              6.490861   0.196831  32.977
## ns(year, df = 4)1:sexMale                         0.319494   1.147815   0.278
## ns(year, df = 4)2:sexMale                         1.179130   1.144024   1.031
## ns(year, df = 4)3:sexMale                         4.477882   2.470508   1.813
## ns(year, df = 4)4:sexMale                         1.067283   0.856289   1.246
## ns(year, df = 4)1:ns(age_mid, df = 4)1           -1.721063   0.573489  -3.001
## ns(year, df = 4)2:ns(age_mid, df = 4)1           -0.128079   0.637631  -0.201
## ns(year, df = 4)3:ns(age_mid, df = 4)1           -2.316199   1.221180  -1.897
## ns(year, df = 4)4:ns(age_mid, df = 4)1            0.717976   0.536268   1.339
## ns(year, df = 4)1:ns(age_mid, df = 4)2           -0.113664   0.411375  -0.276
## ns(year, df = 4)2:ns(age_mid, df = 4)2            1.230501   0.457631   2.689
## ns(year, df = 4)3:ns(age_mid, df = 4)2           -0.539059   0.878617  -0.614
## ns(year, df = 4)4:ns(age_mid, df = 4)2            0.822805   0.384673   2.139
## ns(year, df = 4)1:ns(age_mid, df = 4)3           -1.998860   1.217291  -1.642
## ns(year, df = 4)2:ns(age_mid, df = 4)3            2.851402   1.344503   2.121
## ns(year, df = 4)3:ns(age_mid, df = 4)3           -2.113889   2.584161  -0.818
## ns(year, df = 4)4:ns(age_mid, df = 4)3            2.399539   1.125818   2.131
## ns(year, df = 4)1:ns(age_mid, df = 4)4           -0.777440   0.231379  -3.360
## ns(year, df = 4)2:ns(age_mid, df = 4)4           -0.512831   0.243399  -2.107
## ns(year, df = 4)3:ns(age_mid, df = 4)4           -1.129192   0.485614  -2.325
## ns(year, df = 4)4:ns(age_mid, df = 4)4            0.137824   0.208113   0.662
## sexMale:ns(age_mid, df = 4)1                      2.143142   1.007264   2.128
## sexMale:ns(age_mid, df = 4)2                      2.143267   0.726832   2.949
## sexMale:ns(age_mid, df = 4)3                      5.286729   2.081695   2.540
## sexMale:ns(age_mid, df = 4)4                      1.032512   0.381470   2.707
## ns(year, df = 4)1:sexMale:ns(age_mid, df = 4)1    0.002842   1.114181   0.003
## ns(year, df = 4)2:sexMale:ns(age_mid, df = 4)1   -0.753708   1.108246  -0.680
## ns(year, df = 4)3:sexMale:ns(age_mid, df = 4)1   -3.850881   2.396308  -1.607
## ns(year, df = 4)4:sexMale:ns(age_mid, df = 4)1   -0.671178   0.830804  -0.808
## ns(year, df = 4)1:sexMale:ns(age_mid, df = 4)2   -0.283639   0.801505  -0.354
## ns(year, df = 4)2:sexMale:ns(age_mid, df = 4)2   -1.071936   0.797153  -1.345
## ns(year, df = 4)3:sexMale:ns(age_mid, df = 4)2   -3.249252   1.727587  -1.881
## ns(year, df = 4)4:sexMale:ns(age_mid, df = 4)2   -0.896480   0.596519  -1.503
## ns(year, df = 4)1:sexMale:ns(age_mid, df = 4)3   -0.769264   2.316768  -0.332
## ns(year, df = 4)2:sexMale:ns(age_mid, df = 4)3   -2.195176   2.305693  -0.952
## ns(year, df = 4)3:sexMale:ns(age_mid, df = 4)3   -9.392782   4.967784  -1.891
## ns(year, df = 4)4:sexMale:ns(age_mid, df = 4)3   -2.037017   1.734958  -1.174
## ns(year, df = 4)1:sexMale:ns(age_mid, df = 4)4    0.387372   0.429376   0.902
## ns(year, df = 4)2:sexMale:ns(age_mid, df = 4)4   -0.284371   0.415647  -0.684
## ns(year, df = 4)3:sexMale:ns(age_mid, df = 4)4   -0.934111   0.913947  -1.022
## ns(year, df = 4)4:sexMale:ns(age_mid, df = 4)4   -0.539394   0.320615  -1.682
##                                                 Pr(>|z|)
## (Intercept)                                      < 2e-16 ***
## ns(year, df = 4)1                               0.076796 .
## ns(year, df = 4)2                               0.165411
## ns(year, df = 4)3                               0.185748
## ns(year, df = 4)4                               0.121020
## sexMale                                         0.011974 *
## ns(age_mid, df = 4)1                             < 2e-16 ***
## ns(age_mid, df = 4)2                             < 2e-16 ***
```

```
## ns(age_mid, df = 4)3                             < 2e-16 ***
## ns(age_mid, df = 4)4                             < 2e-16 ***
## ns(year, df = 4)1:sexMale                       0.780744
## ns(year, df = 4)2:sexMale                       0.302688
## ns(year, df = 4)3:sexMale                       0.069904 .
## ns(year, df = 4)4:sexMale                       0.212615
## ns(year, df = 4)1:ns(age_mid, df = 4)1          0.002691 **
## ns(year, df = 4)2:ns(age_mid, df = 4)1          0.840802
## ns(year, df = 4)3:ns(age_mid, df = 4)1          0.057869 .
## ns(year, df = 4)4:ns(age_mid, df = 4)1          0.180623
## ns(year, df = 4)1:ns(age_mid, df = 4)2          0.782315
## ns(year, df = 4)2:ns(age_mid, df = 4)2          0.007170 **
## ns(year, df = 4)3:ns(age_mid, df = 4)2          0.539525
## ns(year, df = 4)4:ns(age_mid, df = 4)2          0.032438 *
## ns(year, df = 4)1:ns(age_mid, df = 4)3          0.100578
## ns(year, df = 4)2:ns(age_mid, df = 4)3          0.033940 *
## ns(year, df = 4)3:ns(age_mid, df = 4)3          0.413347
## ns(year, df = 4)4:ns(age_mid, df = 4)3          0.033058 *
## ns(year, df = 4)1:ns(age_mid, df = 4)4          0.000779 ***
## ns(year, df = 4)2:ns(age_mid, df = 4)4          0.035121 *
## ns(year, df = 4)3:ns(age_mid, df = 4)4          0.020056 *
## ns(year, df = 4)4:ns(age_mid, df = 4)4          0.507808
## sexMale:ns(age_mid, df = 4)1                    0.033363 *
## sexMale:ns(age_mid, df = 4)2                    0.003190 **
## sexMale:ns(age_mid, df = 4)3                    0.011097 *
## sexMale:ns(age_mid, df = 4)4                    0.006796 **
## ns(year, df = 4)1:sexMale:ns(age_mid, df = 4)1 0.997965
## ns(year, df = 4)2:sexMale:ns(age_mid, df = 4)1 0.496447
## ns(year, df = 4)3:sexMale:ns(age_mid, df = 4)1 0.108053
## ns(year, df = 4)4:sexMale:ns(age_mid, df = 4)1 0.419168
## ns(year, df = 4)1:sexMale:ns(age_mid, df = 4)2 0.723426
## ns(year, df = 4)2:sexMale:ns(age_mid, df = 4)2 0.178721
## ns(year, df = 4)3:sexMale:ns(age_mid, df = 4)2 0.059999 .
## ns(year, df = 4)4:sexMale:ns(age_mid, df = 4)2 0.132877
## ns(year, df = 4)1:sexMale:ns(age_mid, df = 4)3 0.739858
## ns(year, df = 4)2:sexMale:ns(age_mid, df = 4)3 0.341063
## ns(year, df = 4)3:sexMale:ns(age_mid, df = 4)3 0.058659 .
## ns(year, df = 4)4:sexMale:ns(age_mid, df = 4)3 0.240354
## ns(year, df = 4)1:sexMale:ns(age_mid, df = 4)4 0.366964
## ns(year, df = 4)2:sexMale:ns(age_mid, df = 4)4 0.493872
## ns(year, df = 4)3:sexMale:ns(age_mid, df = 4)4 0.306752
## ns(year, df = 4)4:sexMale:ns(age_mid, df = 4)4 0.092496 .
## ---
## 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:   3486.4  on 1858  degrees of freedom
## AIC: 12694
##
## Number of Fisher Scoring iterations: 5
```