Biostatistics 1 - Assignment 1

Group 3: William Fleury Dziuk, Galya Gadzhalova, Chen Ma, Afroditi-Iliana Zotou

2025-12-02

Introduction

Within this assignment we will analyse data on incident cases of colon cancer in Sweden, across calendar year as well as by age and sex. The data contain the number of colon cancer cases and demographics of the population in Sweden (age, year and sex on July 1st). The purpose is to describe the incidence of colon cancer in Sweden, especially the incidence pattern across calendar year.

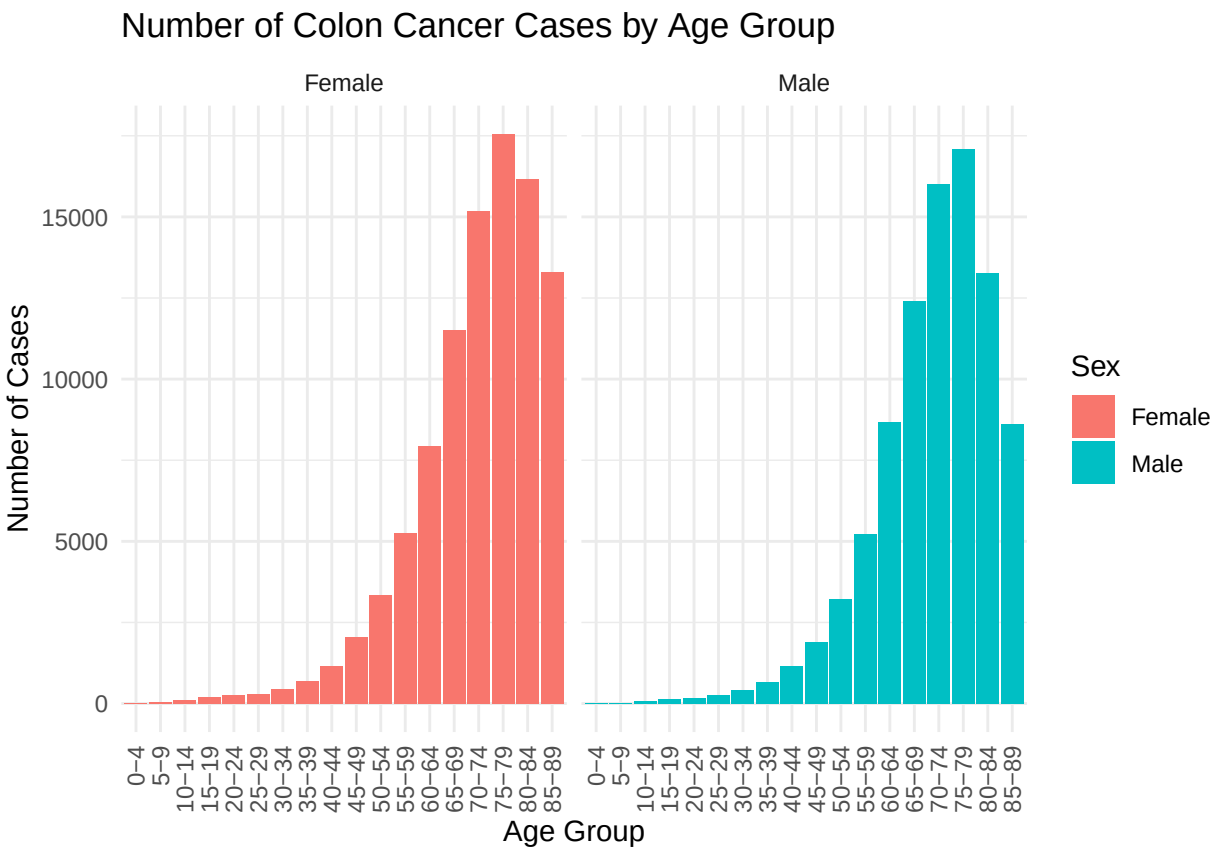
1

Task: Read in the file on number of colon cancer cases (the file cases.tsv) and make sure that you understand the variables included. Create a graph showing the number of cases by age group and sex. Describe what you can conclude from the graph.

Answer:

In our file 'cases.tsv' we have the following variables:

- Age group: Intervals of 5-years (0–4, 5–9, . . . , 85–89).
- Sex: Male and female categories.
- Year: Numeric variable of the calendar year of the observation.
- Number of cases (n): The number of new colon cancer cases in each age group, sex, and year. This is our dependent variable of interest.

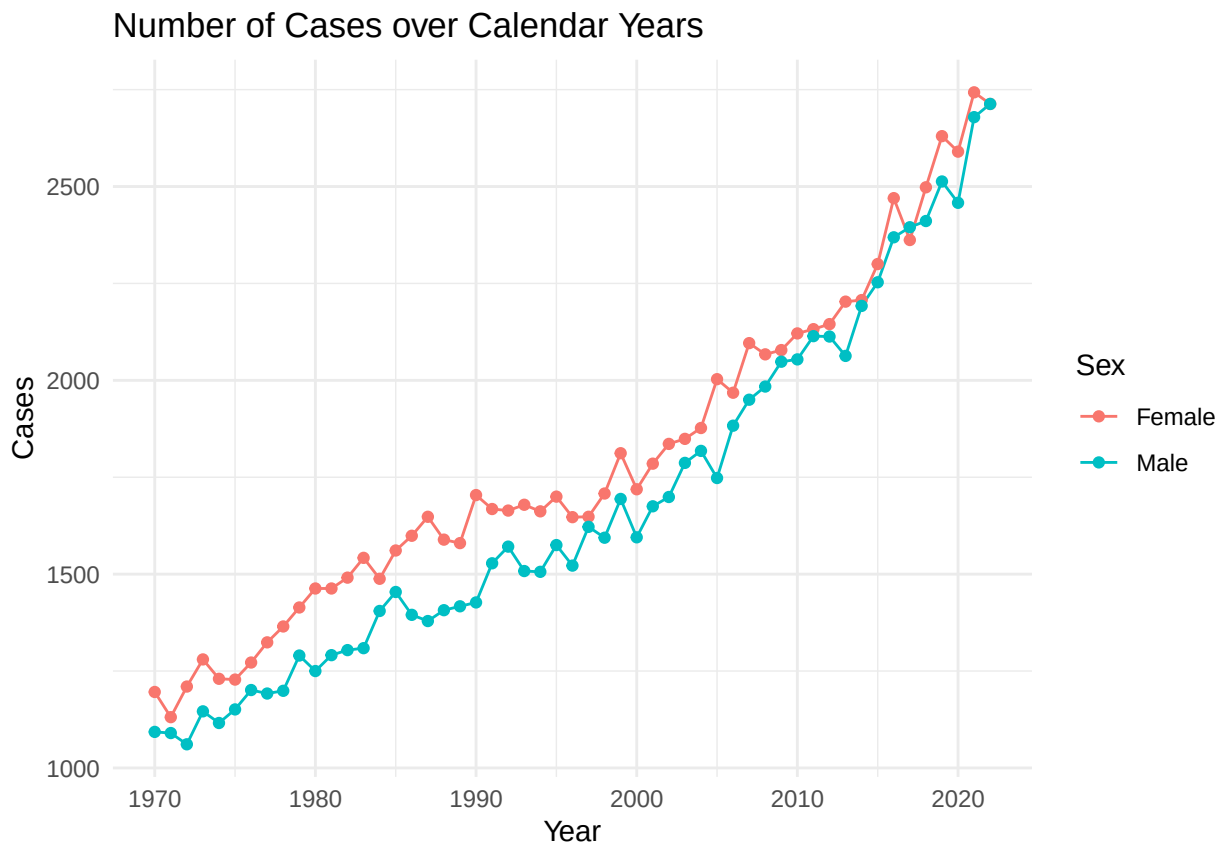


From the graph, we can conclude that the number of cases of colon cancer increases with age for both sexes. Even though there are reported cases in children and young adults of 0-29 years they are rare. A significant increase is shown for both sexes around the 50-54 age group and from that point it dramatically rises in the older age groups. For females, the highest number of cases is seen in the age group 75-79 while for males it is within the age group 80-84 years. This indicates a higher burden of cases in the oldest male populations. We can hence conclude that age is a determining factor in the absolute number of colon cancer cases in Sweden.

2

Task: Obtain the total number of cases in each calendar year by males and females. Create graphs showing the number of cases over calendar years, separately for males and females. Describe what you can conclude from the graphs.

Answer:



The graph shows an overall increasing long-term trend in the incidence of colon cancer in Sweden. There is a steady rise in the absolute number of colon cancer cases for both males and females over the entire 50-year period. Cases for both sexes started around 1.000–1.200 in 1970 and have more than doubled to 2.600–2.700. As for the rate of increase, it was gradual in the period between 1970-2000, but from the 2000 and onward the rise is significantly faster. Throughout the 50-year period the number of cases in females has remained slightly higher compared to males, while the gap between the two lines narrows at the end of our examined period.

3

Task: Read in the file on number of persons at risk (the file population.tsv). Make sure that you understand the variables included. Create graphs that illustrate the population size over age groups and calendar year simultaneously, separately by males and females.

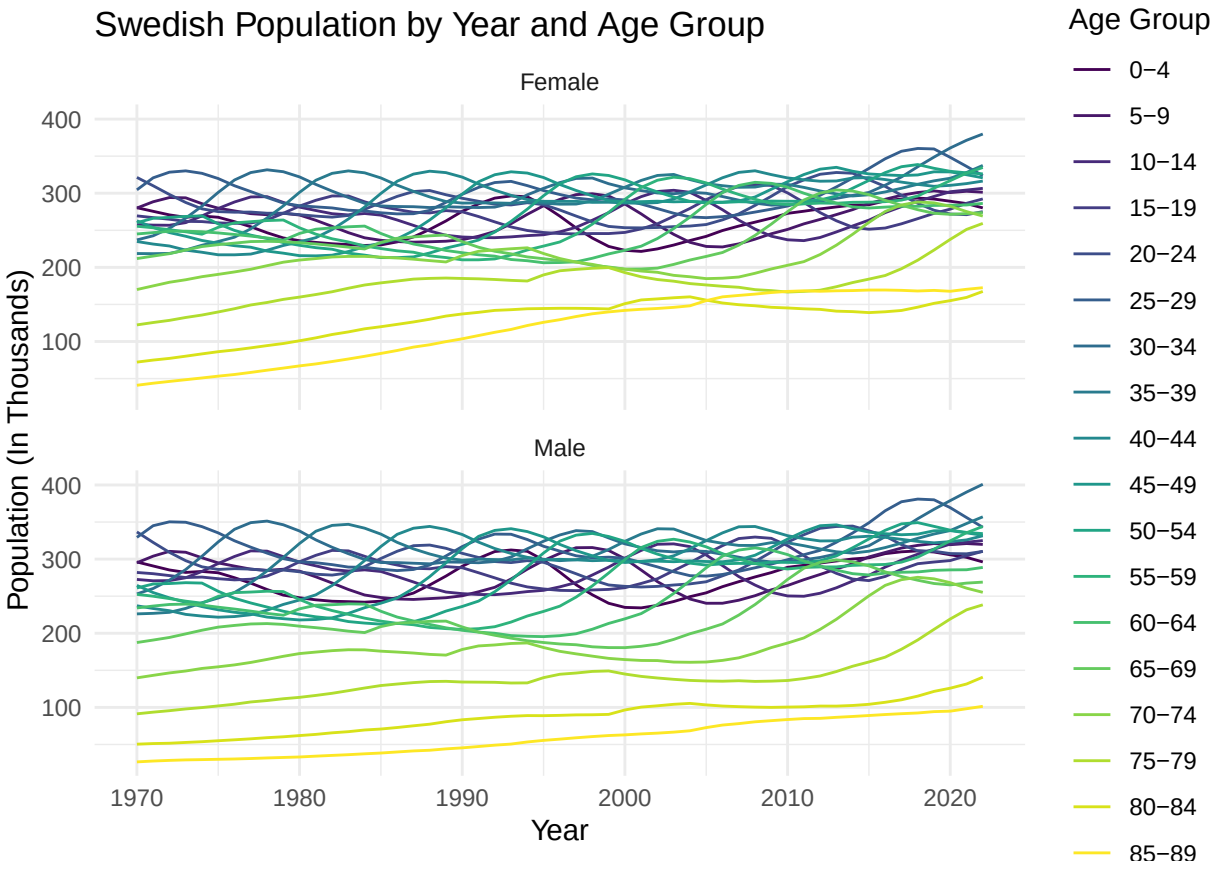
Answer:

```
#Clean up the Dataframe
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"),
  ordered = TRUE
)

population$sex <- factor(population$sex)
```

```r
population$year <- as.numeric(population$year)
population$n_pop <- as.numeric(population$n_pop)
```

In our file 'population.tsv' we have the following variables:

- Age group: Intervals of 5-years (0–4, 5–9, . . . , 85–89).
- Sex: Male and female categories.
- Year: Numeric variable of the calendar year of the observation.
- Number of persons at risk (n_{pop}): The number of individuals in each age group, sex, and year. This represents the population at risk and will be used as the denominator for rate calculations.



These graphs show that the older age groups tend to have a smaller population, which is expected. We can also see that population trends between men and women are similar except for the oldest age groups where there seem to be more women than men, yet again this is expected. The wave like shapes come from the fact that as the years progress, groups of the population transition up to later age groups. We decided to modify the y-axis to "in thousands" because the values are very large otherwise and make it difficult to read the graph.

4

Task: Merge the information on number of cases and the number of persons at risk in each year, for each age group and sex. Does the population file include the same age groups and calendar years as the file

including the number of cases? Also create a separate data frame with the total number of cases and the total population size in each calendar year by males and females.

Answer:

```
#Merged information on number of cases and the number of persons at risk in each year,
#for each age group and sex.
merged_df <- merge(cases, population, by = c("agegroup", "year", "sex"))

#Separate data frame with the total number of cases
#and the total population size in each calendar year by males and females.

cases_sum <- aggregate(n ~ year + sex, data = merged_df, sum)
population_sum <- aggregate(n_pop ~ year + sex, data = merged_df, sum)
summary_df <- merge(cases_sum, population_sum, by = c("year", "sex"))
```

The population file does in fact include the same age groups and calendar years as the file including the number of cases.

5

Task: Create a new variable for the incidence rate of colon cancer by dividing the number of cases with the population size. Do this for both the data including all the age groups and the data with the total number of cases per year and sex. Describe shortly what an incidence rate is, and your thoughts on if this is an appropriate way of calculating an incidence rate.

Answer:

We estimate the Incidence rate (IR) as Incidence Rate $= \frac{\text{Number of new colon cancer cases}}{\text{Population size}}$

We also calculated the rate per 1000 to easier interpret the results. We were motivated to so from similar questions at the recommended exercises.

Incidence Rate per $1000 = 1000 \times \frac{\text{Number of new colon cancer cases}}{\text{Population size}}$

```
##
##  Example rows (summary_df)

##    year    sex    n   n_pop Incidencerate Incidencerate_per1000
## 24 1981   Male 1291 4118622  0.0003134544             0.3134544
## 1  1970 Female 1196 4045318  0.0002956504             0.2956504
## 97 2018 Female 2498 5087747  0.0004909835             0.4909835
## 95 2017 Female 2362 5037580  0.0004688759             0.4688759
## 36 1987   Male 1379 4152583  0.0003320825             0.3320825

##
##  Example rows (merged_df)

##      agegroup year    sex   n  n_pop Incidencerate Incidencerate_per1000
## 1611    75-79 1980 Female 256 159921  1.600790e-03            1.60079039
## 1740    80-84 1991   Male 232  84993  2.729637e-03            2.72963656
## 687     35-39 1995 Female  21 286607  7.327106e-05            0.07327106
## 1053      5-9 2019 Female   0 303099  0.000000e+00            0.00000000
## 741     35-39 2022 Female  11 337476  3.259491e-05            0.03259491
```

From the lecture notes "Introduction to Epidimiology" by Adina Feldman, we define the Incidence rate (IR) as the number of new cases of the outcome divided by the total person-time at risk, for a specific follow-up period. It is a factor used to measure individuals who are newly diagnosed with a disease during a specified period of time.

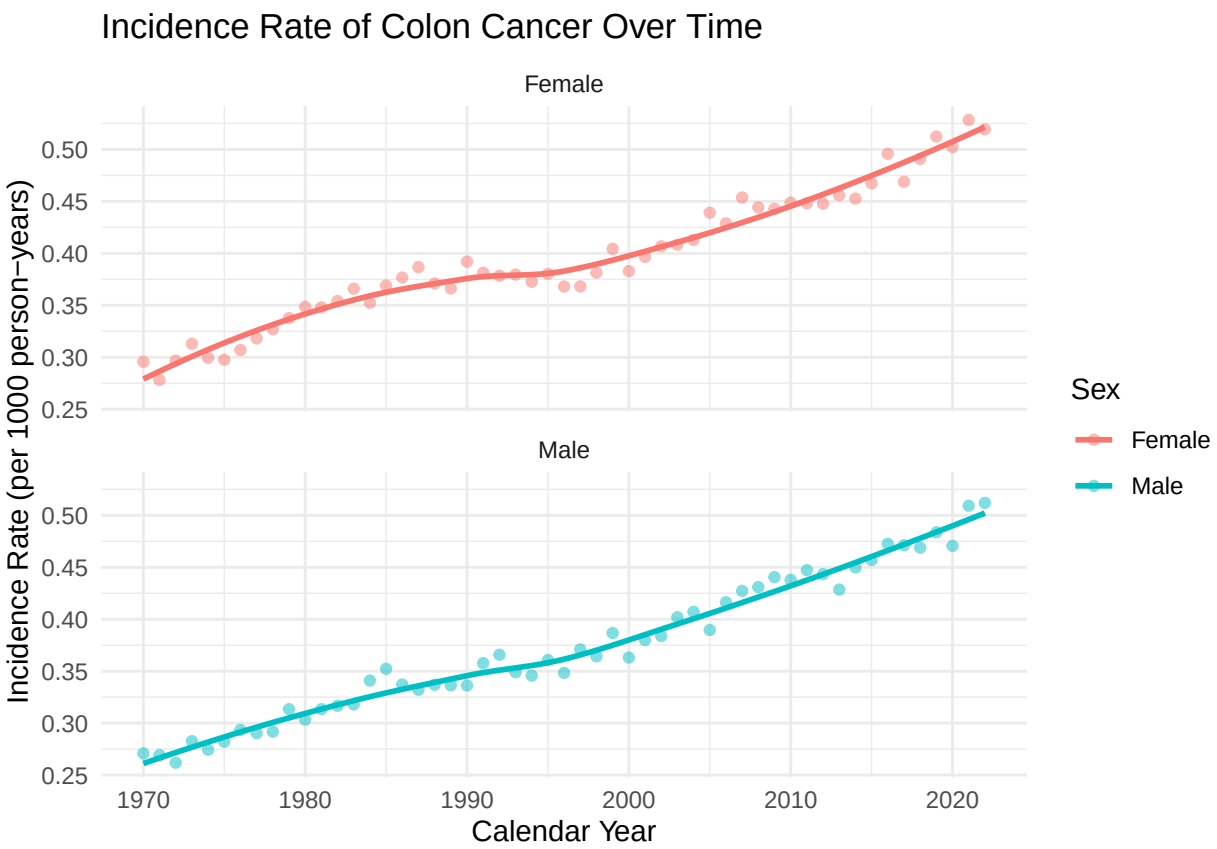
Theoretically, our definition used to calculate the Incidence rate matches Feldman's definition since both measure new cases relative to the population at risk. However, our definition is a simplified version of the formal one, since we assumed the population size to be an approximation for person-time. Our dataset does not include person-time, we approximate the incidence rate by dividing the number of new colon cancer cases by the population size. In our case, the exact person-time is not available, but we can consider that the population does not dramatically change over the year. Hence, we can conclude that this is a valid way of calculating an incidence rate where detailed follow-up is not available.

6

Task: Plot the incidence rate of colon cancer over calendar time and apply a smoother, separately by males and females (do this for the incidence rate based on the total number of cases and the total population size). Describe what you can conclude from the graphs

Answer:

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



From the above graph, we notice that the Incidence Rate is steadily increasing over time for both males and females and it almost doubles in the span of 50 years. This implies that the diagnosis of colon cancer became more common and accessible over the decades. Additionally, the Incidence rates are consistently

higher for males meaning that for the same age period, men have a higher chance of being diagnosed with colon cancer than females. This means that the trends are similar for both sexes but at different absolute levels. As for the LOESS smoother, the curved line, it clearly shows a long-term upward trend by ignoring minor fluctuations.

Confidence bands, as shown by the shaded areas, are relatively narrow, indicating a statistically important trend.

7

Task: Since there is a lot of random variation of the incidence rate from year to year, we can use a regression model to get smooth estimates of the pattern of the incidence rate across calendar year. Fit 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.

Answer:

We will fit a Poisson Regression model for the dependent variable Y that represents the number of new colon cancer cases. We assume: $Y_i \sim \text{Poisson}(\lambda_i)$, where λ_i is the expected number of cases for observation i.

The model links predictors calendar year and sex to the expected count via a log-linear model:

$$\log(\lambda) = \log(\text{population}) + \beta_0 + \beta_1 \cdot \text{year} + \beta_2 \cdot \text{sex}$$

```
# Fit Poisson Regression Model
poisson_model <- glm(
  n ~ year + sex,                  # independent variables
  offset = log(n_pop),             # log of population size as offset
  family = poisson(link = "log"), # Poisson model
  data = summary_df
)

# View model summary
summary(poisson_model)
```

```
##
## Call:
## glm(formula = n ~ year + sex, family = poisson(link = "log"),
##     data = summary_df, offset = log(n_pop))
##
## 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
```

Parameters in the model that can be estimated: β_0, β_1, and β_2.

Estimates that we get: $\beta_0 = -29.69$, $\beta_1 = 0.01094$, $\beta_2 = -0.05592$.

The Poisson regression model is defined as:

$$\log(\lambda) = \log(\text{population}) - 29.69 + 0.01094 \cdot \text{year} - 0.05592 \cdot \text{sex}$$

where λ is the expected number of colon cancer cases and sex is coded as 1 for males and 0 for females.

The value $\beta_0 = -29.69$ represents the baseline log incidence rate for females in year 0.

As for the statistical significance, we see that all coefficients have very small p-values ($< 2 \times 10^{-16}$), indicating strong statistical significance. We can reject the null hypothesis that year and sex have no effect on incidence rate.

The difference between the Null and the Residual deviance suggests that the model explains well the variation in the data. The value AIC is also relatively low which is also an indicator for a good model fit.

The Poisson regression 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 fits the data well and provides a valid estimate of incidence trends across calendar years. It shows that colon cancer incidence rates in Sweden have increased steadily over time and that females have slightly higher rates than males when adjusted for population size.

8

Task: Based on the model output from above, what is the incidence rate in 1970 among males and females? Based on the model output from above, what is the incidence rate in 2020 among males and females? What assumptions have you made regarding how the incidence rate changes over calendar years and what the difference is between males and females?

Answer:

Using the fitted Poisson model in Task 7, the incidence rate is given by:

$$\log(IR) = \beta_0 + \beta_1 \cdot \text{year} + \beta_2 \cdot \text{sex}$$

Therefore, the following calculations can be made:

Incidence rate in 1970:

Females (sex = 0):

$$\log(IR) = -29.69 + 0.01094 \cdot 1970$$

$$\log(IR) = -8.142$$

$$IR_{\text{female, 1970}} = e^{-8.142} = 0.00029$$

Per 1000:

$$IR_{\text{female, 1970}} \times 1000 \approx 0.29$$

Males (sex = 1):

$$\log(IR) = -29.69 + 0.01094 \cdot 1970 - 0.05592$$

then,

$$\log(IR) = -8.198$$

$$IR_{\text{male, 1970}} = e^{-8.198} = 0.000274$$

Per 1000:

$$IR_{\text{male, 1970}} \times 1000 \approx 0.274$$

Similarly, the incidence rate for the year 2020 is calculated, with the only change being that the term $\beta_1 \cdot 1970$ has been replaced with $\beta_1 \cdot 2020$.

A summary table of the model-based incidence rates is presented below:

Year	Sex	Incidence Rate (per 1000)	Incidence Rate
1970	Female	0.29	0.00029
1970	Male	0.27	0.00027
2020	Female	0.52	0.00052
2020	Male	0.49	0.00049

These results indicate that the model predicts a steady increase in incidence over time for both sexes between 1970 and 2020.

The following assumptions could be made based on the model:

- The coefficient for calendar year is $\beta_1 = 0.01094$. When it is normalized, $e^{0.01094} \approx 1.011$, therefore it is concluded that the incidence rate increases by about 1.1% per year. The model assumes that the log incidence rate changes linearly with calendar year. This means that increasing the year by the same amount always changes the incidence rate by the same proportion, no matter which years we start from.
- In the model the variable sex is a binary variable which takes values 0 or 1. So, it is assumed that the relative difference between males and females is constant across all calendar years.

Intrepretation:

The coefficient β_2 for the sex variable is:

$$\beta_2 = -0.05592$$

Then, the value should be normalised, when the exponential is calculated

$$e^{-0.05592} \approx 0.9456$$

This value (0.9456) represents the ratio of the male incidence rate to the female incidence rate.

To find the percentage difference, we subtract 1:

$$0.9456 - 1 = -0.0544$$

which means that males have about 5.4% lower incidence than females.

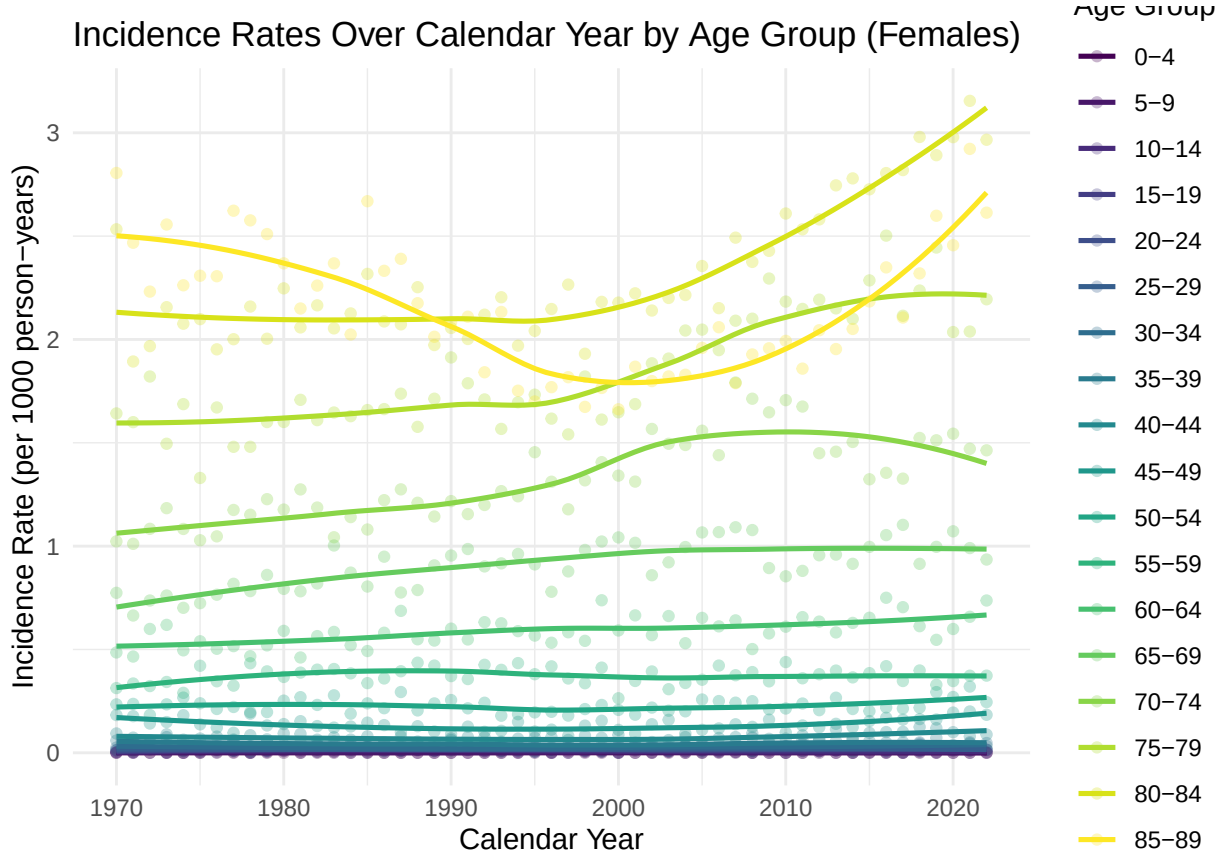
9

Task:

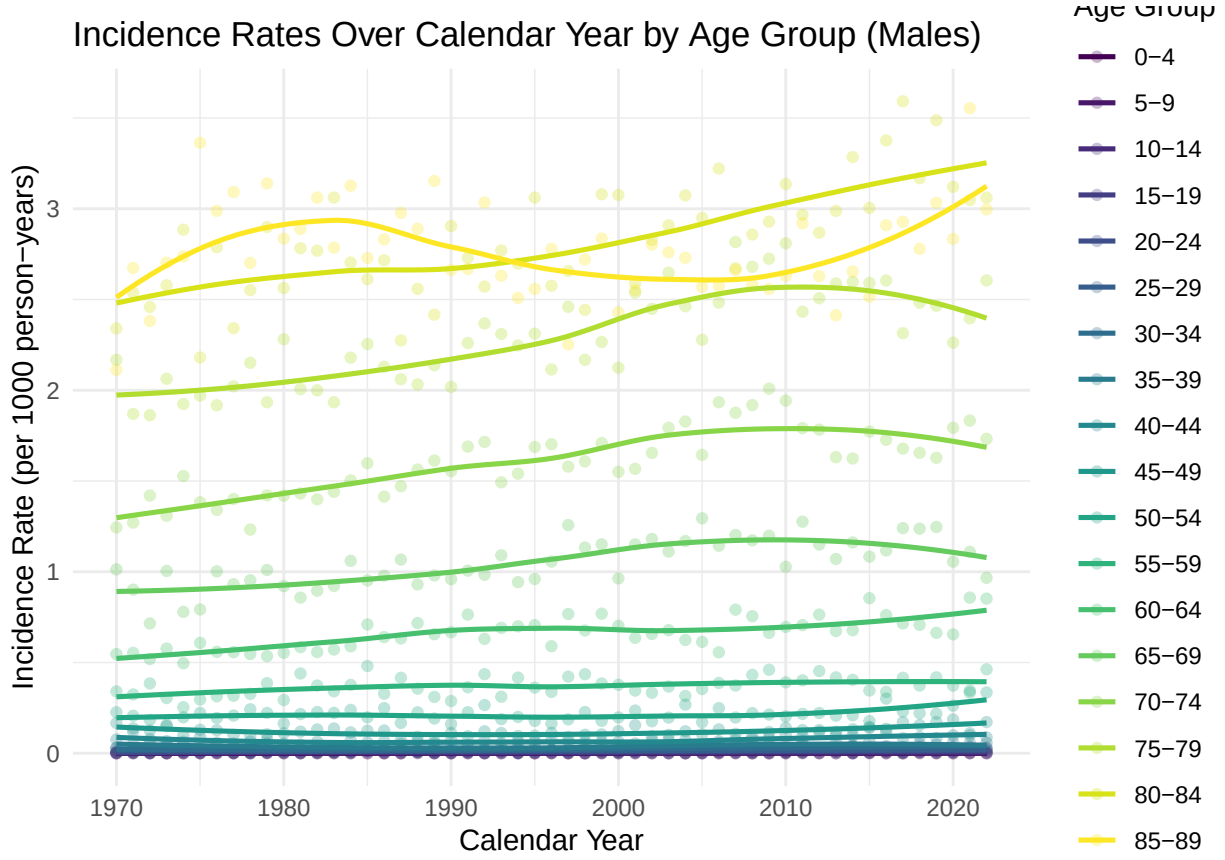
Create a graph of incidence rates over calendar year by sex and age group, and apply smoothers, what can you conclude?

Answer:

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




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



Interpretation of the results:

Across both groups (Female and Male) the incidence rate of colon cancer increases with age. The oldest age group reports the highest incidence rates while the youngest age group reports stable, around zero, rates. Furthermore, the trend over the calendar years is upward, meaning that the incidence rate increases over time especially among the middle-aged groups and older individuals. In contrast younger groups show very little or no change across the years. Overall, the patterns for males and females are similar, with upward trend over time. Also colon cancer incidence rises over time in the older age groups.

10

Task:

Since colon cancer is more common in older age groups, and the age distribution has changed in the population, we want to also adjust for age. Again fit a suitable Poisson model, but this time with the age-specific number of cases as the dependent variable, and age-specific population size as offset, and calendar year, age group and sex as independent variables. Make sure to not assume that the pattern across calendar year is the same for males and females and across age groups. Based on the model output from above, what is the incidence rate in 1970 in age group 70-74 among males and females? Based on the model output from above, what is the incidence rate in 2020 in age group 70-74 among males and females?

Answer:

A suitable model for this task is:

$$\log(\lambda_{i,a,s}) = \log(\text{population}_{i,a,s}) + \beta_0 + \beta_1 \cdot \text{year}_i + \beta_2 \cdot \text{sex}_s + \beta_3 \cdot \text{agegroup}_a + (\text{two-way interactions}).$$

where β_0 is the intercept, i.e. the predicted log-IR for the reference group, β_1 the year variable, β_2 the sex variable and β_3 the age group variable.

Description of the variables in the model:

- `n`: age–sex–year specific number of colon cancer cases (dependent variable).
- 'n_pop': population size in the same age–sex–year group, also `log(n_pop)` as an offset.
- The formula $year + sex + agegroup + year : sex + year : agegroup$ includes all main effects and the two relevant two-way interactions, allowing the time trend to differ by sex and by age group.

```
poisson_age_model_simplier <- glm(
  n ~ year + sex + agegroup +
      year:sex + year:agegroup,
  offset = log(n_pop),
  family = poisson(link = "log"),
  data = merged_df
)
```

```
#Creating new data for predictions for age group 70-74 in 1970 and 2020 from
#the model above.
newdata_70_74 <- data.frame(
  year     = c(1970, 1970, 2020, 2020),
  sex      = factor(c("Female", "Male", "Female", "Male"),
                    levels = levels(merged_df$sex)),
  agegroup = factor(rep("70-74", 4),
                    levels = levels(merged_df$agegroup)),
  #Set n_pop to 1 so predicted value is a rate.
  n_pop    = 1
)

#Predicted incidence rates per person-year.
newdata_70_74$IR <- predict(
  poisson_age_model_simplier,
  newdata = newdata_70_74,
  type = "response"
)

#Convert to incidence per 1000 person-years
newdata_70_74$IR_per_1000 <- newdata_70_74$IR * 1000

newdata_70_74
```

```
##   year    sex agegroup n_pop          IR IR_per_1000
## 1 1970 Female    70-74     1 0.001122754    1.122754
## 2 1970   Male    70-74     1 0.001318093    1.318093
## 3 2020 Female    70-74     1 0.001566451    1.566451
## 4 2020   Male    70-74     1 0.001822175    1.822175
```

The results show that:

- Incidence rates are higher for males than for females in both years.
- Incidence rates are higher in 2020 than in 1970 for both sexes.

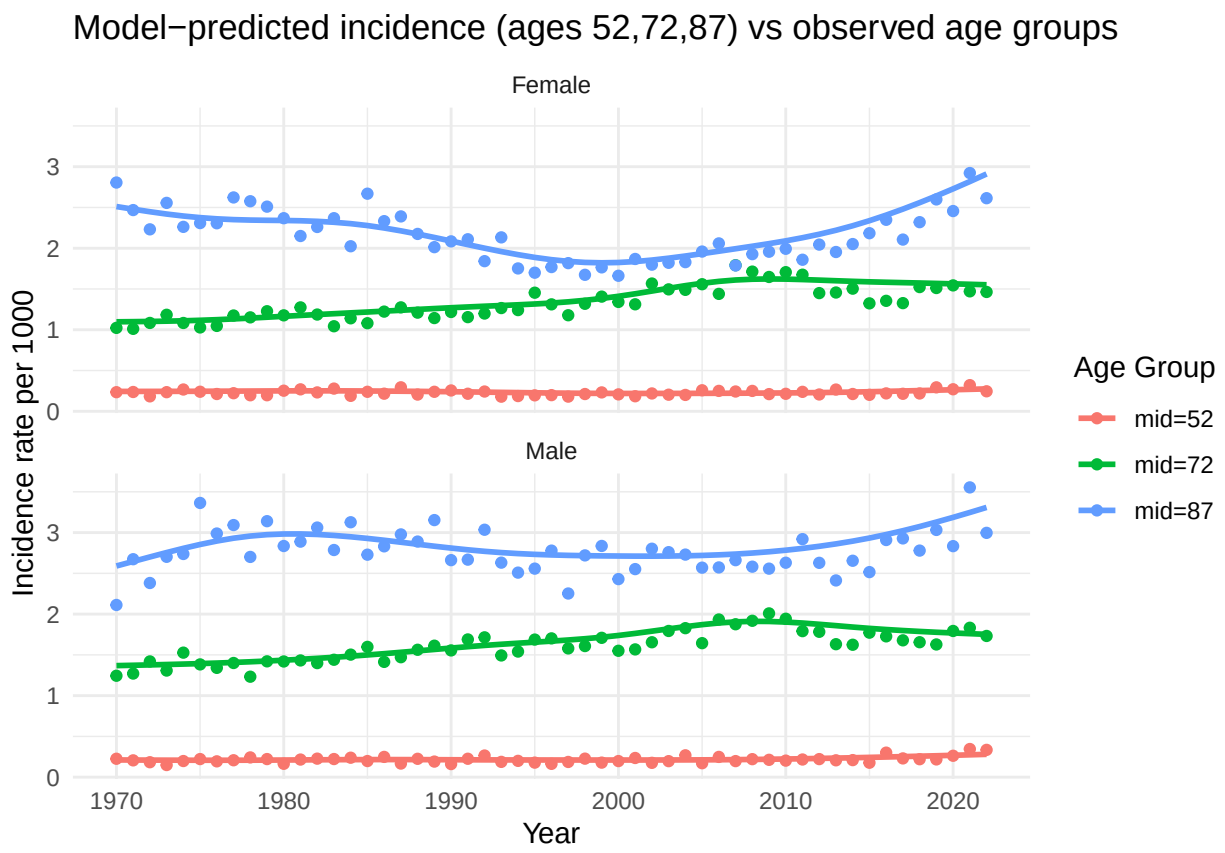
11

Task: If you have not already done so, refit the model above using splines for the effect of calendar year and age group (use the mid point of each age group), and also make sure to not assume that the pattern across calendar year is the same across age and sex. Create graphs showing the incidence rate across calendar time for males and females at ages 52, 72 and 87. Compare with the observed values for age group 50-54, 70-74 and 85-89 from previously.

Answer:

```
if (!"age_mid" %in% names(merged_df)) {
  merged_df <- merged_df %>%
    mutate(
      age_start = as.numeric(sub("^(\\d+)-.*$", "\\1", agegroup)),
      age_end   = as.numeric(sub("^.*-(\\d+)$", "\\1", agegroup)),
      age_mid   = (age_start + age_end) / 2
    ) %>% select(-age_start, -age_end)
}


merged_df$sex <- factor(merged_df$sex)

# Fit GAM: year, age spline
gam_model <- gam(
  n ~ sex +
    s(age_mid,  by = sex, k = 8) +
    s(year,     by = sex, k = 10) +
    ti(year, age_mid, by = sex, k = c(8,6)),
  offset = log(n_pop),
  family = poisson,
  data = merged_df,
  method = "REML"
)


# Predict for ages 52,72,87 across all years
pred_ages <- c(52, 72, 87)
years     <- sort(unique(merged_df$year))
sex_levels <- levels(merged_df$sex)

newpred <- expand.grid(
  year = years,
  sex  = sex_levels,
  age_mid = pred_ages
)

newpred$n_pop <- 1
newpred$pred_rate <- predict(gam_model, newdata = newpred, type = "response")
newpred$pred_rate_per1000 <- newpred$pred_rate * 1000
newpred$age_label <- paste0("mid=", newpred$age_mid)

# Observed: pick groups 50-54, 70-74, 85-89
obs_groups <- c("50-54", "70-74", "85-89")
age_mid_map <- data.frame(
```

```r
  agegroup = obs_groups,
  age_mid = c(52, 72, 87)
)

obs_df <- merged_df %>%
  filter(agegroup %in% obs_groups) %>%
  group_by(year, sex, agegroup) %>%
  summarise(n = sum(n), n_pop = sum(n_pop), .groups = "drop") %>%
  mutate(obs_rate_per1000 = n / n_pop * 1000) %>%
  left_join(age_mid_map, by = "agegroup") %>%
  mutate(age_label = paste0("mid=", age_mid))

# Plot: model (line) vs observed (points)
ggplot() +
  geom_line(
    data = newpred,
    aes(x = year, y = pred_rate_per1000, color = age_label),
    linewidth = 1
  ) +
  geom_point(
    data = obs_df,
    aes(x = year, y = obs_rate_per1000, color = age_label),
    linewidth = 2
  ) +
  facet_wrap(~ sex, ncol = 1) +
  scale_color_discrete(name = "Age Group") +
  labs(
    title = "Model-predicted incidence (ages 52,72,87) vs observed age groups",
    x = "Year", y = "Incidence rate per 1000"
  ) +
  theme_minimal()
```



We fitted a GAM model of the form

$$log(\lambda) = log(n_{\text{pop}}) + \beta_0 + \beta_1 \text{sex} + s(\text{age(mid)}, \text{by} = \text{sex}) + s(\text{year}, \text{by} = \text{sex}) + ti(\text{year}, \text{age(mid)}, \text{by} = \text{sex})$$

which allows age effects, calendar-year trends, and their interactions to differ between males and females. The model shows excellent fit (deviance explained 99.4%).

Compared to the basic Poisson model in Task 7, which assumed a constant log-linear model trend, and the interaction model in Task 10, which allowed for slope differences but constrained them to be linear, the GAM provides superior flexibility. The GAM successfully captures non-linear patterns, which explain almost all the variability in the aggregated data.

We predicted incidence rates at ages 52, 72, and 87 and compared them with the observed incidence in the age groups 50–54, 70–74, and 85–89. Model-based curves closely follow the empirical data.

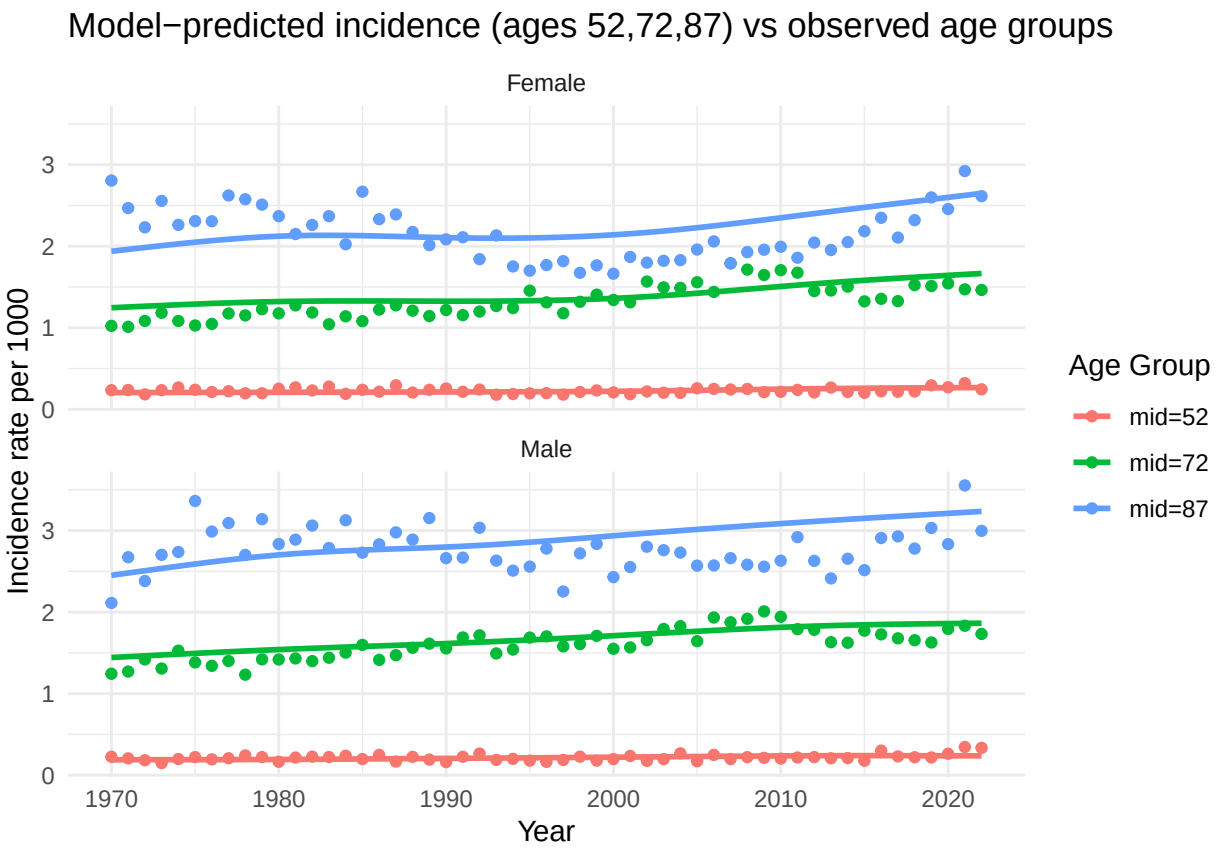
Age 52: Incidence is low ($\approx$ 0.2–0.3 per 1000) and nearly constant over time for both sexes.
Age 72: Incidence increases slowly until the mid-1990s, stabilizes, and shows a slight decline after 2010.
Age 87: Incidence displays a stronger nonlinear pattern: an increase until the 1990s, a dip around 2000–2010, and an increase again in recent years.

Across all ages, males have consistently higher incidence rates than females, although the temporal patterns are similar.
Overall, the spline-based GAM reproduces the observed trends well and appropriately captures the nonlinear effects of age, calendar year, and their interaction.

```r
if (!"age_mid" %in% names(merged_df)) {
  merged_df <- merged_df %>%
    mutate(
      age_start = as.numeric(sub("^(\\d+)-.*$", "\\1", agegroup)),
      age_end   = as.numeric(sub("^.*-(\\d+)$", "\\1", agegroup)),
      age_mid   = (age_start + age_end) / 2
    ) %>% select(-age_start, -age_end)
}


merged_df$sex <- factor(merged_df$sex)


library(splines)

spline_model <- glm(
  formula = n ~ sex +
    ns(year, df = 5) +
    ns(age_mid, df = 6) +
    sex:ns(year, df = 5) +
    sex:ns(age_mid, df = 6) +
    # add interaction between year and age, allow different ages have different trend
    ns(year, df = 5):age_mid,
  family = poisson(link = "log"),
  data = merged_df,
  offset = log(n_pop)
)



# Predict for ages 52,72,87 across all years
pred_ages <- c(52, 72, 87)
years     <- sort(unique(merged_df$year))
sex_levels <- levels(merged_df$sex)

newpred <- expand.grid(
  year = years,
  sex  = sex_levels,
  age_mid = pred_ages
)

newpred$n_pop <- 1
newpred$pred_rate <- predict(spline_model, newdata = newpred, type = "response")
newpred$pred_rate_per1000 <- newpred$pred_rate * 1000
newpred$age_label <- paste0("mid=", newpred$age_mid)

# Observed: pick groups 50-54, 70-74, 85-89
obs_groups <- c("50-54", "70-74", "85-89")
age_mid_map <- data.frame(
  agegroup = obs_groups,
  age_mid = c(52, 72, 87)
)
```

```r
obs_df <- merged_df %>%
  filter(agegroup %in% obs_groups) %>%
  group_by(year, sex, agegroup) %>%
  summarise(n = sum(n), n_pop = sum(n_pop), .groups = "drop") %>%
  mutate(obs_rate_per1000 = n / n_pop * 1000) %>%
  left_join(age_mid_map, by = "agegroup") %>%
  mutate(age_label = paste0("mid=", age_mid))
```

```r
# Plot: model (line) vs observed (points)
ggplot() +
  geom_line(
    data = newpred,
    aes(x = year, y = pred_rate_per1000, color = age_label),
    linewidth = 1
  ) +
  geom_point(
    data = obs_df,
    aes(x = year, y = obs_rate_per1000, color = age_label),
    linewidth = 2
  ) +
  facet_wrap(~ sex, ncol = 1) +
  scale_color_discrete(name = "Age Group") +
  labs(
    title = "Model-predicted incidence (ages 52,72,87) vs observed age groups",
    x = "Year", y = "Incidence rate per 1000"
  ) +
  theme_minimal()
```



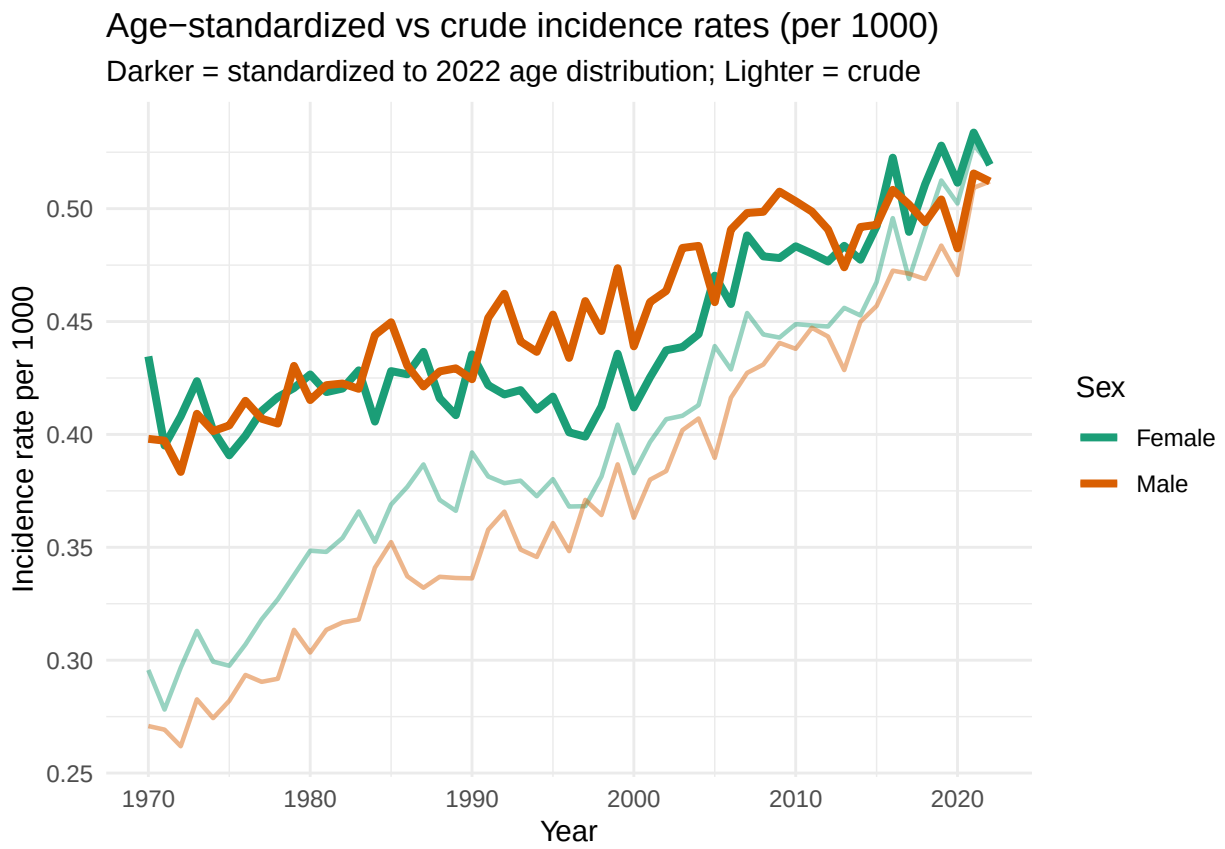
We fitted a Poisson regression model using natural cubic splines (ns) via the splines package. To account for non-linear trends and interactions as requested, we specified the model as follows:

n ~ sex + ns(year, df=5) + ns(age_mid, df=6) + sex:ns(year, df=5) + sex:ns(age_mid, df=6) + ns(year, df=5):age_mid

We increased the degrees of freedom to 5 for year and 6 for age to capture the complex temporal and age-related patterns. The model resulted in a Residual Deviance of 3426.2 and an AIC of 12592, which is a substantial improvement over simpler parametric models and adequately approximates the trends observed in the GAM analysis.

12

Task: If we want to compare incidence rates between calendar years, we typically want to have a summary statistics over all age groups. However, we have to take into account differences in the age distribution between calendar years if we don't want any differences to be due to the population getting older. Age-standardized rates allow us to do this. Estimate direct age standardised incidence rate by year and sex based on the sex-specific age distribution in 2022. Create a graph of age-standardized incidence rates and compare with non-age-standardized graph created previously.



The comparison of incidence rates (IRs) shows that the crude IRs (lighter lines) increase steeply due to the ageing of the Swedish population. In contrast, the age-standardized IRs (thicker lines), which represent the true underlying risk, show a slower increase.

Crucially, while crude IRs sometimes suggest similar risks between sexes in the early years, the standardized IRs demonstrate that Male IR is consistently higher than Female IR across the entire period, confirming that males have a higher age-specific risk.

In general, the non-age-standardized (crude) rates show both the true change in risk, i.e. the incidence rate, and the change in the population's age structure (older people mean higher risk). In our case, we use the 2022 age distribution, so the standardized rates minimize the aging effect and show only the absolute change in incidence risk as if the population's age remained constant (at the 2022 level) throughout the entire period.

13

Task: It is also possible to get age-standardised rates based on the regression model including age, year and sex. Instead of standardising the observed rates, standardisation is applied to the predicted rates from the model. Do so, again using the sex-specific age distribution on 2022. Compare these standardised rates to the direct standardised rates from above.

Answer:

The indirect standardization method uses the predicted age-specific incidence rates (Predicted $\text{IR}_{a,y,s}$) derived from the Poisson regression model (fitted in Task 10) and applies the same 2022 sex-specific age weights ($\text{Weight}_{a,s}^{2022}$) used in Task 12.

The model-based standardized incidence rate is calculated as:

$$\text{Model-based Standardized IR}_{\text{y,s}} = \sum_{a} \left(\text{Predicted IR}_{a,y,s}^{\text{model}} \times \text{Weight}_{a,s}^{2022} \right)$$

```
##   year    sex          ASR ASR_per_1000
## 1 1970 Female 0.0003815197    0.3815197
## 2 2020 Female 0.0005012038    0.5012038
## 3 1970   Male 0.0003970554    0.3970554
## 4 2020   Male 0.0005086264    0.5086264
```

Interpretation:

The model-based ASR increases from 1970 to 2020 in both sexes, even after age adjustment, indicating a true temporal rise in colon cancer incidence.

Males consistently have slightly higher age-standardised rates than females.

Compared to the directly standardised rates, the model-based ASR are smoother and less affected by random variation, because the regression model removes year-to-year noise.

Both approaches show the same overall pattern: rising incidence over time and slightly higher rates in males.

14

Task: What do you conclude regarding the pattern of colon cancer incidence across calendar years?

Answer:

Overall, we can conclude that the incidence of colon cancer in Sweden has risen steadily from 1970 to 2020 for both sexes. The smoothed plot of the yearly incidence rates (Task 6) shows an almost doubling of the rate over the 50-year span with an upward trend. The Poisson regression model (Task 7) quantifies this rise as an average 1.1 increase per calendar year ($\beta_1 = 0.01094$).

Even after adjusting for the ageing population (Task 12), the age-standardised incidence curves still rise, less sharply than the crude rates but still verifying that the trend reflects a valid increase. Throughout the period males consistently have higher rates than females, but the general pattern is similar for both sexes.

Hence, our data indicate a nationwide increase in colon cancer incidence over the last five decades affecting equally men and women. In a more practical perspective, we can interpret that there have been major improvements in screening and diagnosing colon cancer over the past decades and awareness has been raised encouraging testing especially for men.