Assignment 1

2024-07-01

#### Question 1

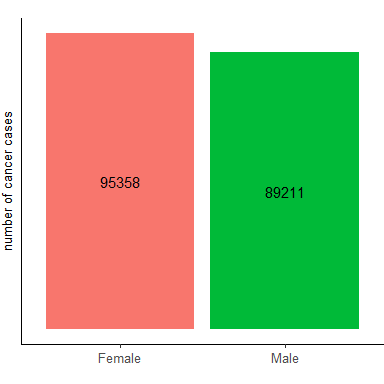
Read in the data file on number of colon cancer cases (the file *cases.RData*) and make sure that you understand the variables included.

# Read in R data set  
  
load(paste0(path,"cases.RData"))  
  
# Look at the data set:  
View(cases)  
  
# Look at descriptive statistics for the data set:  
describe(cases)  
  
# Look at the first rows of the data set:  
head(cases)  
  
# Create a codebook for the data set (not necessary for this exercise)  
  
# codebook(cases,survey\_overview = FALSE,  
# detailed\_scales = FALSE,  
# detailed\_variables = FALSE,  
# missingness\_report = FALSE,  
# metadata\_json = FALSE)

**Create graphs showing the number of cases by sex and by age groups.**

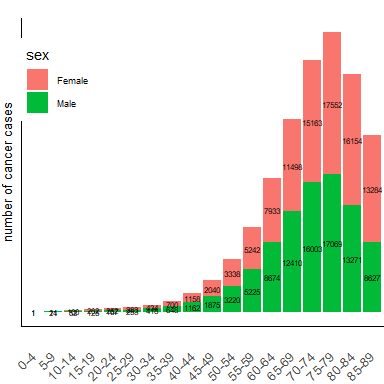
Describe what you can conclude from the graphs.

# Aggregate the data set by sex   
totals <- cases %>% group\_by(sex) %>%   
 summarize(value = sum(n))  
  
cases\_sex <- ggplot(totals, aes(x = sex, y = value,fill = sex)) +   
 geom\_col() +   
 # define a general theme for the graph  
 theme\_classic() +  
 # make some changes in the theme  
 theme(  
 legend.position = "none",  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"), #define margins for the plot  
 axis.title.x = element\_blank(), # take away x-axis title   
 axis.text.x = element\_text(size = 10), # define size for x-axis values   
 axis.title.y = element\_text(size = 10), # define size for y-axis title  
 axis.text.y = element\_blank(), # take away y-axis values   
 axis.ticks.y = element\_blank()) + # take away y-axis ticks   
 labs(y = "number of cancer cases")+ # define y-axis title  
 scale\_fill\_manual(values=cols) + # assign manually colors to the graph  
 # include y-values inside the bars  
 geom\_text(data = totals, aes(x = sex, y = value, label = value),position = position\_stack(vjust = 0.5))  
  
cases\_sex



**Create graphs by sex and by age groups**

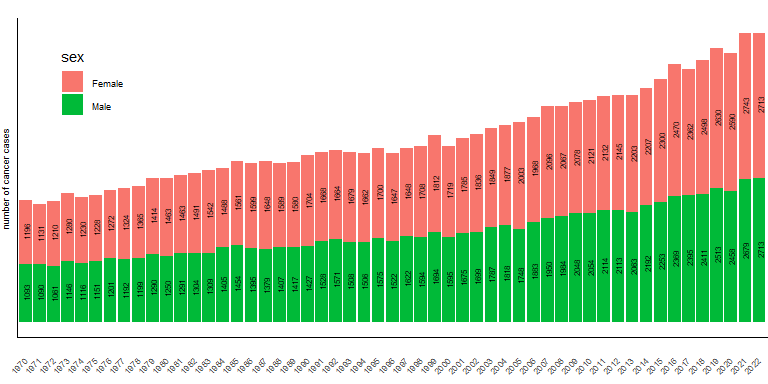
# Calculate # of cases by sex and age group  
totals\_age <- cases %>%  
 group\_by(sex,agegroup) %>%  
 summarize(value = sum(n))  
  
cases\_age <- ggplot(totals\_age, aes(x = agegroup, y = value, fill = sex)) +   
 geom\_col() +   
 theme\_classic() +  
 theme(  
 legend.text = element\_text(size=7),  
 legend.position = "inside",  
 legend.position.inside = c(0.1,0.8),  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"),  
 axis.title.x = element\_blank(),  
 axis.text.x = element\_text(size = 10, angle = 45,vjust=0.3, hjust=1),  
 axis.ticks.x = element\_blank(),  
 axis.title.y = element\_text(size = 10),  
 axis.text.y = element\_blank(),  
 axis.ticks.y = element\_blank())+  
 labs(x = "5-year age groups",y = "number of cancer cases")+   
 scale\_fill\_manual(values=cols) +  
 geom\_text(data = totals\_age, aes(x = agegroup, group = sex,y = value, label=value),position = position\_stack(vjust = 0.5),size = 2)  
  
cases\_age



#### Question 2

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

# Calculate # of cases by sex and calendar year  
totals\_year <- cases %>%  
 group\_by(sex, year) %>%  
 summarize(value = sum(n))  
  
  
cases\_year <- ggplot(totals\_year, aes(x = as.factor(year), y = value, fill = sex)) +   
 geom\_col() +   
 theme\_classic() +  
 theme(legend.text = element\_text(size=7))+  
 theme(  
 legend.position = "inside",  
 legend.position.inside = c(0.1,0.8),  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"),  
 axis.title.x = element\_blank(),  
 axis.text.x = element\_text(size = 7,family = "Arial", angle = 45,vjust=0.3, hjust=1),  
 axis.title.y = element\_text(size = 7,family = "Arial"),  
 axis.text.y = element\_blank(),  
 axis.ticks.y = element\_blank(),  
 axis.ticks.x = element\_blank())+  
 labs(x = "5-year age groups",y = "number of cancer cases")+   
 scale\_fill\_manual(values=cols) +  
 geom\_text(aes(group = sex, label=value),position = position\_stack(vjust = 0.5),size = 2,angle = 90)  
  
cases\_year



#### Question 3:

Merge on information on number of persons at risk (the file *population.RData*) in each year, for each age group and sex. Make sure that you understand the variables included. Does the population file include the same age groups and calendar years as the file including the number of cases? Why can this be important to check?

# Read in population data set:  
  
load(paste0(path,"population.RData"))  
describe(population)  
head(population)

# Merge two data sets on sex, age group and year:  
joined\_df <- left\_join(cases,population)  
  
head(joined\_df)

## # A tibble: 6 × 5  
## agegroup year sex n n\_pop  
## <fct> <dbl> <fct> <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

#### Question 4:

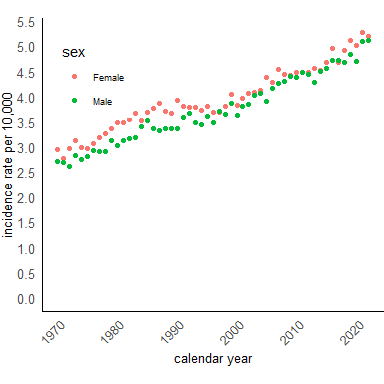
Create a new variable for the incidence rate of colon cancer by dividing the number of cases with the population size. Describe shortly what an incidence rate is, and your thoughts on if this is an appropriate way of calculating an incidence rate.

joined\_df <- joined\_df %>% mutate(ir = n / n\_pop,   
 ir\_extra = ir\*10000)

#### Question 5:

Collapse the joined data set by sex and calendar year. Create a ne variable of incidence rate. Plot the incidence rate of colon cancer over calendar time, separately by males and females. Describe what you can conclude from the graphs.

# Create a variable of total # of cancer cases by calendar year and sex  
joined\_df\_collapsed <- joined\_df %>% group\_by(sex,year) %>%   
 summarise(total\_pop = sum(n\_pop), total\_cases = sum(n))  
  
# Create IR per 1 person & per 10,000 people:  
joined\_df\_collapsed <- joined\_df\_collapsed %>%   
 mutate(ir = total\_cases /total\_pop) %>%   
 mutate(ir\_extra = ir\*10000)  
  
ir\_year <- ggplot(joined\_df\_collapsed, aes(x = year, y = ir\_extra, color = sex)) +   
 geom\_point() +   
 theme\_classic() +  
 theme(  
 legend.text = element\_text(size=7),  
 legend.position = "inside",  
 legend.position.inside = c(0.15,0.8),  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"),  
 axis.title.x = element\_text(size = 10),  
 axis.text.x = element\_text(size = 10, angle = 45, hjust=1),  
 axis.ticks.x = element\_blank(),  
 axis.title.y = element\_text(size = 10),  
 axis.text.y = element\_text(size = 10),  
 axis.ticks.y = element\_blank())+  
 scale\_y\_continuous(limits = c(0,NA),n.breaks = 10)+  
 labs(x = "calendar year",y = "incidence rate per 10,000") +   
 scale\_color\_manual(values=cols)   
  
ir\_year

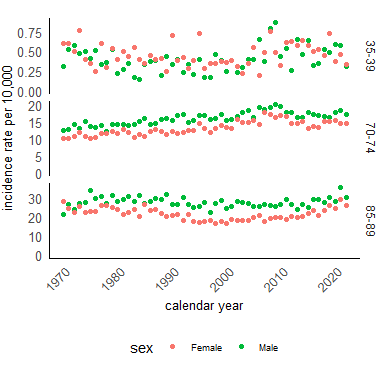


# There is an increasing incidence rate for men and women across years

#### Extra:

Create a graph of incidence rates by calendar year, sex for age groups [35-39], [70-74] & [84-89]

temp <- joined\_df %>% filter(agegroup == "35-39" | agegroup == "70-74" | agegroup == "85-89")  
  
ir\_year\_age <- ggplot(temp, aes(x = year, y = ir\_extra, color = sex)) +   
 geom\_point() +   
 theme\_classic() +  
 theme(  
 legend.text = element\_text(size=7),  
 legend.position = "bottom",  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"),  
 axis.title.x = element\_text(size = 10),  
 axis.text.x = element\_text(size = 10, angle = 45, hjust=1),  
 axis.ticks.x = element\_blank(),  
 axis.title.y = element\_text(size = 10),  
 axis.text.y = element\_text(size = 10),  
 axis.ticks.y = element\_blank())+  
 scale\_y\_continuous(limits = c(0,NA),n.breaks = 5)+  
 labs(x = "calendar year",y = "incidence rate per 10,000") +   
 scale\_color\_manual(values=cols)   
  
  
plot <- ir\_year\_age + facet\_grid(rows = vars(agegroup), scales = "free") +  
 theme(  
 strip.background = element\_blank(),  
 strip.text.x = element\_text(  
 size = 7, face = "bold.italic")  
 )  
  
plot



#### Question 6:

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 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.

# Create a binary variable female with values 0/1, where female = 0 if sex == F   
  
joined\_df\_collapsed <- joined\_df\_collapsed %>% mutate(female = ifelse(sex == "Female", 0,1))  
  
  
m1 <- glm(total\_cases ~ female + year, family="poisson"(link = "log"), offset = log(total\_pop), data=joined\_df\_collapsed)  
# Look at the output:  
m1

##   
## Call: glm(formula = total\_cases ~ female + year, family = poisson(link = "log"),   
## data = joined\_df\_collapsed, offset = log(total\_pop))  
##   
## Coefficients:  
## (Intercept) female year   
## -29.69415 -0.05592 0.01094   
##   
## Degrees of Freedom: 105 Total (i.e. Null); 103 Residual  
## Null Deviance: 5505   
## Residual Deviance: 222.8 AIC: 1211

#summary(m1)

#### Question 7:

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?

# Create a newdata set for predictions by subsetting joined\_df\_total:  
  
newdata1 <- joined\_df\_collapsed %>% filter(year == 1970 |year == 2020) %>%   
 mutate(total\_pop = 1)   
  
# Obtain predicted IR per person   
newdata1$pred\_ir <- predict(m1, newdata1, type = "response")  
  
head(newdata1[,c(1,2,5,8)])

## # A tibble: 4 × 4  
## # Groups: sex [2]  
## sex year ir pred\_ir  
## <fct> <dbl> <dbl> <dbl>  
## 1 Female 1970 0.000296 0.000293  
## 2 Female 2020 0.000502 0.000507  
## 3 Male 1970 0.000271 0.000277  
## 4 Male 2020 0.000471 0.000479

# Obtain predicted IR per 10 000:  
newdata1 <- newdata1 %>% mutate(total\_pop = 10000)  
newdata1$pred\_ir\_1 <- predict(m1, newdata1, type = "response")  
  
# type = response is an equivalent to exp(predict)  
  
head(newdata1[,c(1,2,6,9)])

## # A tibble: 4 × 4  
## # Groups: sex [2]  
## sex year ir\_extra pred\_ir\_1  
## <fct> <dbl> <dbl> <dbl>  
## 1 Female 1970 2.96 2.93  
## 2 Female 2020 5.02 5.07  
## 3 Male 1970 2.71 2.77  
## 4 Male 2020 4.71 4.79

#### Question 8:

Since colon cancer is more common in older age groups, and the age distribution has changed in the population, we want to estimate age-specific rates over calendar years, by sex.

Again fit a 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 age groups and across calendar year is the same for males and females.**

joined\_df <- joined\_df %>% mutate(female = ifelse(sex == "Female", 0,1))  
  
m2 <- glm(n ~ female + year + agegroup + female\*agegroup + female\*year, family="poisson"(link = "log"), offset = log(n\_pop), data=joined\_df)  
  
m2

##   
## Call: glm(formula = n ~ female + year + agegroup + female \* agegroup +   
## female \* year, family = poisson(link = "log"), data = joined\_df,   
## offset = log(n\_pop))  
##   
## Coefficients:  
## (Intercept) female year   
## -27.564288 0.753505 0.005568   
## agegroup5-9 agegroup10-14 agegroup15-19   
## 3.156674 4.578716 5.271921   
## agegroup20-24 agegroup25-29 agegroup30-34   
## 5.447842 5.504617 5.929360   
## agegroup35-39 agegroup40-44 agegroup45-49   
## 6.430484 6.951372 7.533343   
## agegroup50-54 agegroup55-59 agegroup60-64   
## 8.049684 8.541196 9.003359   
## agegroup65-69 agegroup70-74 agegroup75-79   
## 9.436690 9.816248 10.149417   
## agegroup80-84 agegroup85-89 female:agegroup5-9   
## 10.374125 10.252880 -0.779506   
## female:agegroup10-14 female:agegroup15-19 female:agegroup20-24   
## -0.653420 -0.483461 -0.410167   
## female:agegroup25-29 female:agegroup30-34 female:agegroup35-39   
## -0.107884 -0.034137 -0.068972   
## female:agegroup40-44 female:agegroup45-49 female:agegroup50-54   
## 0.020765 -0.057781 0.001239   
## female:agegroup55-59 female:agegroup60-64 female:agegroup65-69   
## 0.050446 0.169257 0.200099   
## female:agegroup70-74 female:agegroup75-79 female:agegroup80-84   
## 0.245340 0.267669 0.254752   
## female:agegroup85-89 female:year   
## 0.325747 -0.000404   
##   
## Degrees of Freedom: 1907 Total (i.e. Null); 1870 Residual  
## Null Deviance: 407000   
## Residual Deviance: 2949 AIC: 12130

#summary(m2)

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?

newdata2 <- joined\_df %>% filter(year == 1970 |year == 2020) %>%   
 filter(agegroup == "70-74" | agegroup == "35-39") %>%   
 mutate(n\_pop = 10000) %>% arrange(sex, agegroup,year)   
  
newdata2 <- joined\_df %>% filter(year == 1970 |year == 2020) %>%   
 mutate(n\_pop = 10000) %>% arrange(sex, agegroup,year)   
  
  
# Obtain predicted IR per person   
newdata2$pred\_ir <- predict(m2, newdata2, type = "response")  
  
head(newdata2[,c(1,2,3,7,9)])

## # A tibble: 6 × 5  
## agegroup year sex ir\_extra pred\_ir  
## <fct> <dbl> <fct> <dbl> <dbl>  
## 1 0-4 1970 Female 0 0.000621  
## 2 0-4 2020 Female 0 0.000820  
## 3 5-9 1970 Female 0 0.0146   
## 4 5-9 2020 Female 0 0.0193   
## 5 10-14 1970 Female 0.0388 0.0605   
## 6 10-14 2020 Female 0.0993 0.0799

#### Question 9:

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.

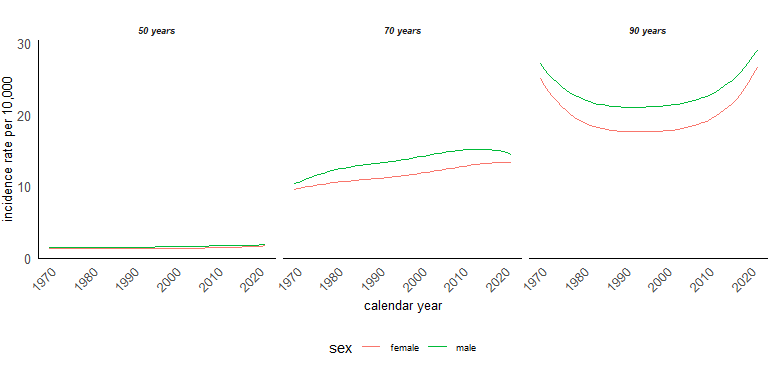
# Create a mid point for each age group:  
joined\_df$age\_mid <- age\_mid(joined\_df$agegroup)  
  
m3 <- glm(n ~ female + nsk(year,df=2)+nsk(age\_mid,df=2) + female\*nsk(year,df=2) + nsk(year,df=2)\*nsk(age\_mid,df=2), family="poisson"(link = "log"), offset = log(n\_pop), data=joined\_df)  
  
m3

##   
## Call: glm(formula = n ~ female + nsk(year, df = 2) + nsk(age\_mid, df = 2) +   
## female \* nsk(year, df = 2) + nsk(year, df = 2) \* nsk(age\_mid,   
## df = 2), family = poisson(link = "log"), data = joined\_df,   
## offset = log(n\_pop))  
##   
## Coefficients:  
## (Intercept)   
## -12.63087   
## female   
## 0.07636   
## nsk(year, df = 2)1   
## -1.40872   
## nsk(year, df = 2)2   
## -0.42757   
## nsk(age\_mid, df = 2)1   
## 3.25541   
## nsk(age\_mid, df = 2)2   
## 6.62986   
## female:nsk(year, df = 2)1   
## 0.10348   
## female:nsk(year, df = 2)2   
## 0.01153   
## nsk(year, df = 2)1:nsk(age\_mid, df = 2)1   
## 1.30896   
## nsk(year, df = 2)2:nsk(age\_mid, df = 2)1   
## 0.63354   
## nsk(year, df = 2)1:nsk(age\_mid, df = 2)2   
## 1.17967   
## nsk(year, df = 2)2:nsk(age\_mid, df = 2)2   
## 0.54535   
##   
## Degrees of Freedom: 1907 Total (i.e. Null); 1896 Residual  
## Null Deviance: 407000   
## Residual Deviance: 4081 AIC: 13210

#summary(m3)

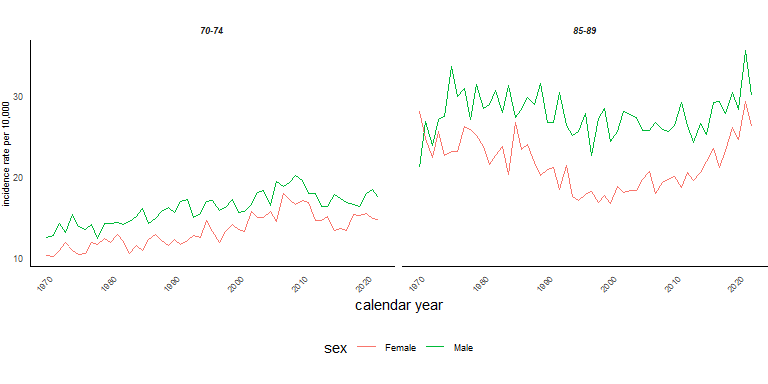
Create graphs showing the incidence rate across calendar time for males and females at ages 50, 70 and 90.

# Create a new data set for predictions  
newdata3 <- tibble(year = rep(c(1970:2022),6),   
 female=c(rep(c(rep(0,53),rep(1,53)),3)),  
 age\_mid = c(rep(50,106),rep(70,106),rep(90,106)),  
 n\_pop = rep(10000,318))  
  
# Obtain predicted IR per 10,000 person   
newdata3$pred\_ir <- predict(m3, newdata3, type = "response")  
  
  
newdata3$female <- factor(newdata3$female, levels = c(0,1), labels = c("female","male"))  
newdata3$age\_mid <- factor(newdata3$age\_mid, levels=c(50,70,90),labels = c("50 years","70 years","90 years"))  
  
plot <- ggplot(data = newdata3, aes(x = year, y = pred\_ir, color = female)) +  
 geom\_line() +   
 theme\_classic() +  
 theme(  
 legend.text = element\_text(size=7),  
 legend.position = "bottom",  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"),  
 axis.title.x = element\_text(size = 10),  
 axis.text.x = element\_text(size = 10, angle = 45, hjust=1),  
 axis.title.y = element\_text(size = 10),  
 axis.text.y = element\_text(size = 10),  
 axis.ticks.y = element\_blank(),  
 axis.ticks.x = element\_blank())+  
 labs(x = "calendar year",y = "incidence rate per 10,000", color = "sex")+   
 scale\_color\_manual(values=cols)   
  
plot <- plot + facet\_grid(cols = vars(age\_mid)) +  
 theme(  
 strip.background = element\_blank(),  
 strip.text.x = element\_text(  
 size = 7, face = "bold.italic"))  
  
  
plot



**Extra** Compare with the observed values for age group [70-74] & [84-89]

temp <- joined\_df %>% select(sex,year,female, ir,agegroup) %>% mutate(ir = ir\*10000) %>% filter(agegroup == "70-74" | agegroup == "85-89")  
  
  
plot <- ggplot(data = temp, aes(x = year, y = ir, color = sex)) +  
 geom\_line() +   
 theme\_classic() +  
 theme(  
 legend.text = element\_text(size=7),  
 legend.position = "bottom",  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"),  
 axis.title.x = element\_text(size = ),  
 axis.text.x = element\_text(size = 7, angle = 45, hjust=1),  
 axis.title.y = element\_text(size = 7),  
 axis.text.y = element\_text(size = 7),  
 axis.ticks.y = element\_blank(),  
 axis.ticks.x = element\_blank())+  
 labs(x = "calendar year",y = "incidence rate per 10,000", color = "sex")+   
 scale\_color\_manual(values=cols)   
  
plot <- plot + facet\_grid(cols = vars(agegroup)) +  
 theme(  
 strip.background = element\_blank(),  
 strip.text.x = element\_text(  
 size = 7, face = "bold.italic")  
 )  
  
plot



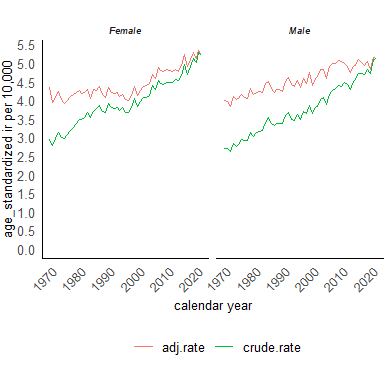
#### EXTRA:

If we want to compare incidence rates between calendar years, we would like to have a summary statistics over all age groups. However, we have to take into account differences in the age distribution between calendar years. Age-standardized rates allow us to make such a comparison. We, first, calculate direct age-standardized incidence rates by taking for “standard” age distribution in the last available calendar year, 2022.

library(epitools)  
library(tidyverse)  
  
# Save a standard population, i.e. the number of population in each age group in year 2022:  
  
standard <- population %>% filter(year == 2022)  
  
# Join our data set with standard values   
  
temp <- left\_join(joined\_df, standard, by = c("agegroup", "sex"), copy = TRUE)  
  
# Calculate direct age-stndardized incidence rates for each calendar year and separately for males and females   
dsr <- temp %>%   
 group\_by(sex,year.x) %>%   
 summarise(age\_adjust = list(ageadjust.direct(count = n,  
 pop = n\_pop.x, stdpop = n\_pop.y))) %>%   
 mutate(age\_adjust = map(age\_adjust, as.data.frame.list)) %>%   
 unnest

Create a graph of age-standardized incidence rates and compare with non-age-standardized graph:

# Change the data set in the long format:  
  
df <- pivot\_longer(dsr,cols = ends\_with("rate"), names\_to = "group",values\_to = "ir") %>%   
 mutate(ir\_extra = ir\*10000)  
  
dsr <- ggplot(df, aes(x = year.x, y = ir\_extra, color = group)) +   
 geom\_line() +   
 theme\_classic() +  
 theme(  
 legend.text = element\_text(size = 10),  
 legend.position = "bottom",  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"),  
 axis.title.x = element\_text(size = 10),  
 axis.text.x = element\_text(size = 10, angle = 45, hjust=1),  
 axis.ticks.x = element\_blank(),  
 axis.title.y = element\_text(size = 10),  
 axis.text.y = element\_text(size = 10),  
 axis.ticks.y = element\_blank())+  
 scale\_y\_continuous(limits = c(0,NA),n.breaks = 10)+  
 labs(x = "calendar year",y = "age\_standardized ir per 10,000", color = "") +   
 scale\_color\_manual(values=cols)   
  
dsr <- dsr + facet\_grid(cols = vars(sex)) +  
 theme(  
 strip.background = element\_blank(),  
 strip.text.x = element\_text(  
 size = 7, face = "bold.italic")  
 )  
  
dsr



Obtain regression standardized incidence rates, i.e. averaged over age:

# Obtain predictions for all age\_mid:  
  
stand <- joined\_df %>% mutate(n\_pop=10000)  
  
stand$pred\_ir\_1 <- predict(m3,stand,type = "response")   
  
# Calculate an average pred\_ir\_1 by sex and calendar year:  
stand <- stand %>% group\_by(sex,year) %>%   
 summarise(pred\_ir\_1\_mean = mean(pred\_ir\_1))

Create graphs of standardized over age for each calendar year and sex:

age\_stand <- ggplot(stand, aes(x = year, y = pred\_ir\_1\_mean, color = sex)) +   
 geom\_line() +   
 theme\_classic() +  
 theme(  
 legend.text = element\_text(size = 10),  
 legend.position = "bottom",  
 plot.margin = unit(c(0.5,0,0.5,0),"cm"),  
 axis.title.x = element\_text(size = 10),  
 axis.text.x = element\_text(size = 10, angle = 45, hjust=1),  
 axis.ticks.x = element\_blank(),  
 axis.title.y = element\_text(size = 10),  
 axis.text.y = element\_text(size = 10),  
 axis.ticks.y = element\_blank())+  
 scale\_y\_continuous(limits = c(0,NA),n.breaks = 10)+  
 labs(x = "calendar year",y = "age\_standardized ir per 10,000", color = "") +   
 scale\_color\_manual(values=cols)   
  
age\_stand

