Assignment 1

Biostatistics 1:Introduction to biostatistics, 5BD000

November 2024

Introduction

Within this assignment you will analyse data on incident cases of colon cancer in Sweden, across calendar year as well as by age and sex. The data is found in 2 files, a file containing the number of colon cancer cases and a file containing the population size of Sweden, by age, year and sex (the population size on July 1st). The purpose is to describe the incidence of colon cancer in Sweden, especially the incidence pattern across calendar year.

Answer the questions below and submit the report as a pdf-file, including your code and the created graphs, on Canvas by xxxx. For this assignment you will work in groups of 3-4 individuals, and each group hands in 1 report.

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. Create a graph showing the number of cases by age group and sex. Describe what you can conclude from the graph.

2. Create a separate dataset with the total number of cases in each calendar year by males and females. Create graphs showing the number of cases over calendar years for males and females separately. Describe what you can conclude from the graphs.

3. Read in the data file on number of persons at risk (the file population.RData). 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?

4. Merge the information on number of cases and the number of persons at risk in each year, for each age group and sex (so merge cases.RData and population.RData). Create a separate dataset for the total population size in each calendar year by males and females, and merge (on year and sex) with the corresponding dataset for cases.

5. In both merged datasets, 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.

6. Plot the incidence rate of colon cancer over calendar time, separately by males and females (here you can use the incidence rate based on the total number of cases and the total population size). Describe what you can conclude from the graphs. Create a graph of incidence rates by calendar year and sex for age groups 50-54, 70-74 and 85-89, what can you conclude?

7. 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.

8. 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?

9. 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. 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?

10. 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 above.

11. If we want to compare incidence rates between calendar years, we might 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. Age-standardized rates allow us to make such a comparison. We, first, calculate direct age-standardized incidence rates by taking for "standard" the age distribution in the last available calendar year, 2022. Estimate direct age-standardised incidence rate by year and sex by averaging over age groups, applying weights for each age group based on the age-distribution in 2022. Create a graph of age-standardized incidence rates and compare with non-age-standardized graph.

12. It is also possible to get age-standardised rates based on the regression model including age, year and sex by the use of regression standardisation, i.e. by averaging over age. Compare these standardised rates to the direct standardised rates.

13. What do you conclude regarding the pattern of colon cancer incidence across calendar years?