MODERN EPIDEMIOLOGY

ASSIGNMENT 1

Hephaes Chuen Chau (M116 I54105029)

2022-02-25

Question 1 (Exercise of Chapter 1)

Question 1(a): After observing the incidence rates by age at any given year, it is concluded that "aging is not re1ated to an increase in the incidence of Y and may even be re1ated to a decrease in the incidence." Do you agree with this observation? Justify your answer.

I do not agree with this observation. If we identify the diagonal incidence rates in the table, we can see that the rates actually increase as calendar time is elapsed for a particular cohort. For example, for the cohort 1930-1934 (corresponding to those aged 20-24 in 1950), the rates increase as time progresses (10, 17, 20, . . .). **This suggests that a strong cohort effect is present despite the fact that** while a calendar time is fixed (eg, 1970), the rates are not distributed as such that higher age predicts a higher rate.

Alternatively, the same data as shown by the table is plotted with the x-axis being the birth cohort and the connected data points representing the data corresponding to a particular cohort. In such a plot, the specificity of the cohort effects is more apparent.

To do this we first define the function for transforming the structure of the data:

```
transform_age_vs_prev <- function(df) {

  result <- list()
  row_length <- nrow(df)
  col_length <- ncol(df)

  # The following is a control sequence for populating a new list

  for (i in seq_len(row_length)) {

    for (j in 2:col_length) {

      vec <- numeric(4)
      vec[2] <- as.numeric(df[i,1]) # This is the age
      vec[3] <- as.integer(colnames(df[i,j])) # This is the period
      vec[1] <- vec[3] - vec[2] # This is the cohort
      vec[4] <- as.numeric(df[i,j]) # This is the prevalence

      result <- list.append(result, vec)

    }

  }

  new_df <- do.call(rbind,result)
  new_df <- data.frame(new_df)

  colnames(new_df) <- c("Cohort", "Age", "Period", "Prevalence")

  return(new_df)

}
```

Before we have the data with the structure as:

```
library(readr)
library(rlist)
df <- read_csv("data_assignment1.csv")
```

```
## Rows: 8 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (9): Age, 1950, 1955, 1960, 1965, 1970, 1975, 1980, 1985
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(df)
```

```
## # A tibble: 6 x 9
##     Age `1950` `1955` `1960` `1965` `1970` `1975` `1980` `1985`
##   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1    22     10     15     22     30     33     37     41     44
## 2    27      8     17     20     24     29     38     40     43
## 3    32      5     12     22     25     28     35     42     45
## 4    37      3     12     15     26     30     32     39     42
## 5    42      2     10     17     19     28     32     39     42
## 6    47      2     12     15     18     21     33     40     42
```

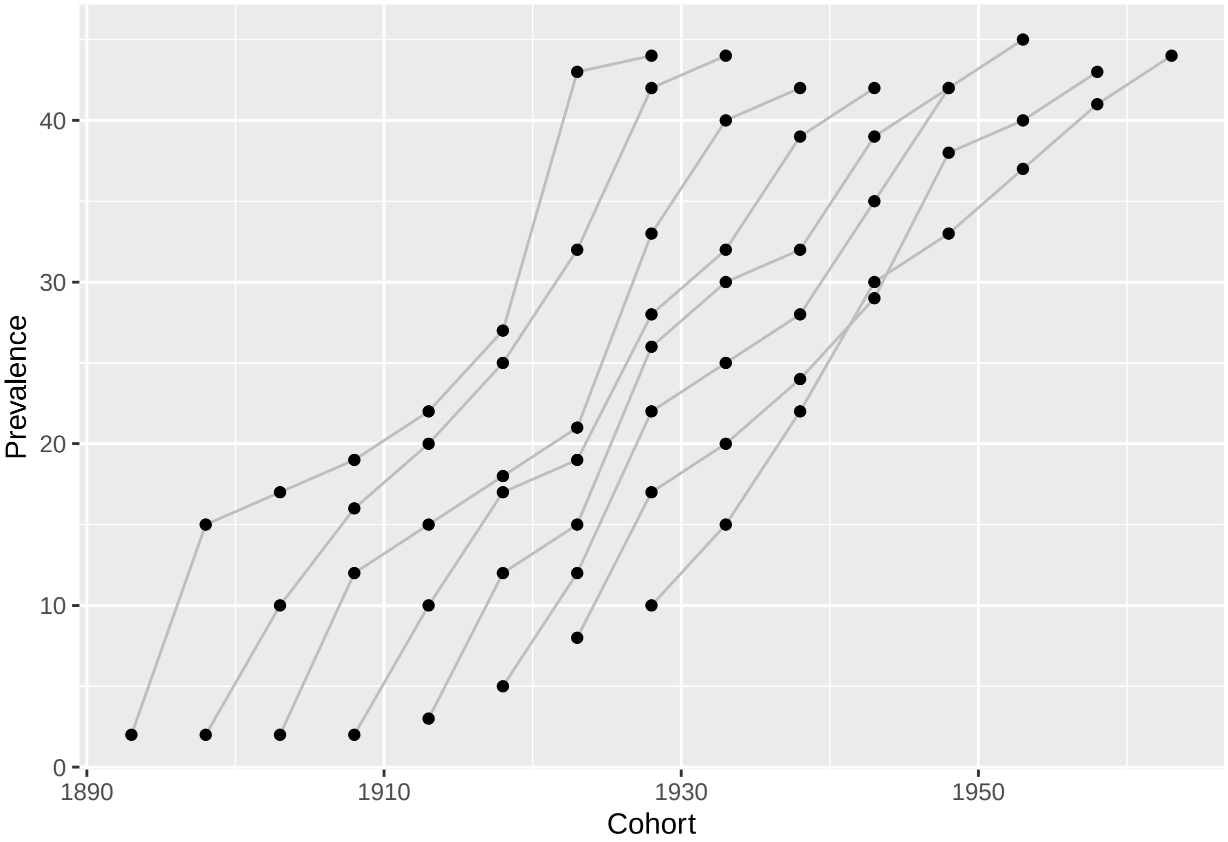
Now the data is transformed as

```
new_df <- transform_age_vs_prev(df)
head(new_df)
```

```
##   Cohort Age Period Prevalence
## 1   1928  22   1950         10
## 2   1933  22   1955         15
## 3   1938  22   1960         22
## 4   1943  22   1965         30
## 5   1948  22   1970         33
## 6   1953  22   1975         37
```

We can thus plot the data (with each line corresponding to the same age group across different cohorts)

```
library(ggplot2)
library(ggrepel)
ggplot(new_df, aes(x = Cohort, y = Prevalence,
                 group = Age)) +
    geom_line(col = 'gray') +
    geom_point()
```



Question 1(b): What are the purposes of examining birth cohort vis-à-vis cross-sectiona1 rates?

Examining the incidence rates of a disease across birth cohorts help us uncover “cohort effects”. **If cohort effects exist, the investigators are led to conjecture what types of life experiences that accumulated and that were characteristic of a particular birth cohort may be responsible for the observed change in incidence rates.** For a hypothetical example, if we identified that the birth cohort in 1930s have increased rates of depression relative to younger cohorts, the investigators may conjecture that war experiences may play a role in affecting the susceptibility of the individuals in the cohort to depression (or that the war experiences may have acted as selective pressure in favour of those prone to depression).

Secondly, **examining the cross-sectional rates reveals the differential burden of the disease for different age groups as the disease exists today.** This may be significant to formulating a correct and cost-effective intervention. Again as a hypothetical example, the knowledge that younger people (as opposed to the elderly) exhibit higher rates of depression may warrant policy interventions that are very likely of different types than what would be expected if it were the elderly that were more depressed.

Question 2: A case-control study is conducted within a well-defined cohort. The reason for this is that expensive additional data collection is needed, and the budget is not sufficient to obtain these data from all cohort participants.

Question 2(a): What type of case-control study within this cohort would be ideal to study multiple outcomes, and why is the alternative case-control design not recommended?

Case-cohort study is ideal for studying multiple outcomes, while nested case-control study is not. There are at least two reasons nested case-control study is not appropriate:

1. In nested case-control study, control is identified simultaneously whenever a case occurs, and the control is identified from the risk set, which is the set of all subjects in the cohort who are at-risk for the event of interest (meaning that they are not the case). **Thus, if our interest is not restricted to only one outcome, it is necessary to draw more than one control group from the cohort.** Since the control groups differ for different outcomes, we cannot infer odds ratio with respect to a factor between the group with two outcomes co-occuring and the control group, because there is not such a control. For case-cohort, since a subcohort is fixed and defined as the control regardless of the outcome being studied, it is possible to draw such an inference.

2. More importantly, as informed by Kim et al, nested-case control study is analysed most frequently using conditional logistic regression model, which is a statistical method that cannot handle the response variable being non-binary (Kim, 2014). Although the same study cited has developed a statistical method to allow for comparison of secondary outcomes in nested-case control study, such method is not the standard and require assumptions stated in the study.

Meanwhile, in case-cohort study, the control is selected as a subcohort (random sample of the total cohort) at baseline, and therefore we can reuse the same control for another outcome of interest.

Question 2(b): In this cohort study, prevalent cases were not excluded at baseline, and thus, the investigators chose to use baseline data to examine associations between suspected risk factors andl preva1ent disease. What type of approach is this, and what are its main advantage and disadvantage?

The approach is case-cohort study. Advantages include:

1. As said in 2(a), in case-cohort study we can reuse the subcohort selected as control to serve as control for multiple cases. An example is provided in the book as ARIC cohort study by Dekker et al.

2. Since the control selected is a sub-cohort, and we have probed the risk factors of this subcohort, we can infer the distributions of risk factors for the cohort.

The disadvantages are:

1. Since the control was not immediately identified as the cases occured (ie, the control is not time-indexed), and they are identified before the cases, we cannot match the cases with subjects without the event at the time the cases occured, unlike in nested-case control study.

2. It is possible the exposure of the cases cannot be reliably ascertained because of recall bias, particularly when the exposure of interest include environmental variables.

3. When the exposure variable of interest potentially changes over time, since we do not re-examine the exposure of the control at a later time in case-cohort study, the estimate of the effects of the exposure could be biased.