Quantifying the Health Impacts of Air Pollution

Day 2: Estimating the risks associated with air pollution

In this practical we will use `R` to perform meta-analyses. We will work through an example by calculating a pooled effect estimate of the excess risk from an increase in PM$_{2.5}$ exposure. All data required for this practical can be found in the folder `Data`. For this practical, you will need the following files

- Relative risks associated with an increase of $10\mu gm^{-3}$ in PM$_{2.5}$ (`AirPollutionStudies.csv`)
- The article from where the example has been extracted - Hoek *et al.* (2013) 'Long-term air pollution exposure and cardio-respiratory mortality: a review' (`Hoeketal2013.pdf`)

Preliminaries

We need the following package

- `metafor` - Package that provides methods for meta-analysis.

As in Practical 1, we use the `install.packages()` function to download and install the packages that we need.

```r
# Installing required packages
install.packages('metafor')
```

We use the `library()` function to load them into the `R` library.

```r
# Loading required packages into the library
library(metafor)
```

Before reading in any data for this practical you will need to ensure that you are in the correct folder. As explained in Practical 1, you can use the `setwd()` function

```r
setwd("Chosen_Directory_Path")
```

If you cannot get the `setwd()` to work, go to `Session > Set Working Directory > Choose Directory` in the toolbar on the top.

Remember, more information about any of the functions used here can be found by typing `help(function_name)` or `?function_name` into `R`.

Example: Risks associated with increases in PM$_{2.5}$

In yesterday's practical, we used a relative risk of 1.06 per $10\mu gm^{-3}$ increase in PM$_{2.5}$ exposure to calculate the number of deaths associated with PM$_{2.5}$. How did we choose 1.06? There are many studies which have estimates of the increased risk of PM$_{2.5}$ exposure, the population at risk. For example, the Netherlands Cohort Study produced a relative risk of 1.06 per $10\mu gm^{-3}$ and the Rome cohort study produced a relative risk of 1.04 per $10\mu gm^{-3}$. Which is the most appropriate to use? Can we use this to calculate the number of deaths associated with PM$_{2.5}$ in Mexico City?

Instead of choosing a value from a particular study, we used a combined (or 'pooled') estimate which arose from a meta analysis. We will use this example, which is taken from the article 'Long-term air pollution exposure and cardio-respiratory mortality: a review' by Hoek *et al.* (2013), to perform a meta analysis. This paper is a systematic review which looks at summarising the long-term risks associated with increases in air pollution. We want to calculate a pooled effect estimate of the excess risk per $10\mu gm^{-3}$ increase in $PM_{2.5}$ exposure.

There are three steps associated with a meta-analysis:

- Extracting
  - main results from each study considered, for example, Relative Risks, Odds Ratios
  - estimates of whether the result may have occurred by chance, for example, Standard Errors, Confidence Intervals.

- Checking
  - whether it is appropriate to calculate a pooled summary/average result across the studies
  - appropriateness depends on just how different the individual studies are that you are trying to combine.
- Calculation
  - by summarising results as a weighted average across the studies using a specified model.

Extracting and Checking

Table 1 in Hoek *et al.* (2013) contains a summary of relative risk estimates (excess risk per 10 μgm^{-3}) from all cohort studies on particulate matter (PM_{10} or $PM_{2.5}$) mortality from all causes and cardiovascular diseases.

Activities

- Look at the studies in this table. Do you think all studies should be included in this meta analysis? If not, which ones and why?

We only include 11 of the studies listed in Table 1 of Hoek *et al.* (2013). `AirPollutionStudies.csv` contains log relative risks associated with an increase of $10\mu gm^{-3}$ in $PM_{2.5}$ and a measure of uncertainty around these log relative risks. These are in csv format, so we use the `read.csv()` function to read them into `R`.

```r
# Reading in log relative risks from separate studies
RR_bystudy <- read.csv('AirPollutionStudies.csv')
```

To check that the data has been read into `R` correctly, we can print the dataset to view its contents

```r
# Printing the dataset
RR_bystudy
                           Study         beta          se
1   American Cancer Society (18)  0.058268908 0.021570762
2  Netherlands Cohort Study (23)  0.058268908 0.045632452
3             Nurses Health (25)  0.231111721 0.105096885
4      Health Professionals (29) -0.154688509 0.088815094
5               US truckers (32)  0.095646780 0.036207462
6           ACS Los Angeles (19)  0.157003749 0.054483189
7           Canadian cohort (34)  0.095310180 0.023207086
8       California teachers (36)  0.009950331 0.035069130
9           Medicare cohort (26)  0.043059489 0.006106133
10              Rome cohort (36)  0.039220713 0.004905960
11                 Six city (16)  0.131028262 0.033467401
```

We can see that this dataset has 11 studies and contains the following variables:

- `Study` - Study name,
- `beta` - Log relative risk associated with an increase in air pollution,
- `se` - Standard error estimate of log relative risk.

We can also summarise the dataset using the `summary()` function. This will allow us to check for anomalies in our data.

```
# Summarising the data
summary(RR_bystudy)
                          Study         beta                 se
 ACS Los Angeles (19)        :1   Min.   :-0.15469   Min.   :0.004906
 American Cancer Society (18):1   1st Qu.: 0.04114   1st Qu.:0.022389
 California teachers (36)    :1   Median : 0.05827   Median :0.035069
 Canadian cohort (34)        :1   Mean   : 0.06947   Mean   :0.041324
 Health Professionals (29)   :1   3rd Qu.: 0.11334   3rd Qu.:0.050058
 Medicare cohort (26)        :1   Max.   : 0.23111   Max.   :0.105097
 (Other)                     :5
```

Activities

- Does it look like `R` has read in the data correctly?
- The data above is the log relative risks. Create a new column which shows the relative risks of an increased effect of $PM_{2.5}$
- Do all studies say there is an increased risk of air pollution? If not, which one? Why do you think this is the case?

Calculation

Now the results have been extracted and verified, a model will need to be fit to find an over all result. This can be done by fitting fixed or random effect models to the studies.

Fixed effect models assume all of the studies examined are considered to have been conducted under similar conditions with similar subjects. They assume the only difference between studies is their power to detect the outcome of interest and assumes there is a single 'true' or 'fixed' underlying effect. They can be used where there is no evidence of heterogeneity.

Random effect models assume that the true treatment/exposure effects in the individual studies may be different from each other. They allow the study outcomes to vary in a normal distribution between studies and assumes there is no single effect to estimate but a distribution of effects due to between-study variation. They tend to give more conservative results than fixed effects as it includes an extra source of variation (between study).

Activities

- Refer back to Table 1 in Hoek *et al.* (2013), do you think we should fit a fixed or random effects model? Why?

The function `rma()` allows us to perform a meta analysis, to create a pooled effect estimate of the excess risk per $10\mu gm^{-3}$ increase in $PM_{2.5}$ exposure as discussed in the lecture yesterday and today.

```
# Fitting the random effects model
mod <- rma(yi  = beta, # Log RR estimates
           sei = se, # Standard error estimates
           slab = Study, # Study name for labelling
           data = RR_bystudy, # Dataset
           method = 'DL') # Method: FE for Fixed Effects and DL for Random Effects
```

All the information from our meta-analysis is contained in the object `mod`. This can be summarised using the `summary()` function.

```
# Summarising meta-analysis
summary(mod)
```

```
Random-Effects Model (k = 11; tau^2 estimator: DL)

  logLik  deviance       AIC       BIC      AICc
 15.0242   27.2404  -26.0483  -25.2525  -24.5483

tau^2 (estimated amount of total heterogeneity): 0.0004 (SE = 0.0005)
tau (square root of estimated tau^2 value):      0.0207
I^2 (total heterogeneity / total variability):   65.04%
H^2 (total variability / sampling variability):  2.86

Test for Heterogeneity:
Q(df = 10) = 28.6016, p-val = 0.0014

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.0606  0.0104  5.8247  <.0001  0.0402  0.0810  ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model summary contains the result of the pooled estimate as well and many assessments of the meta analysis including Cochran's Q and the I^2 statistic.

Cochran's Q calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies.

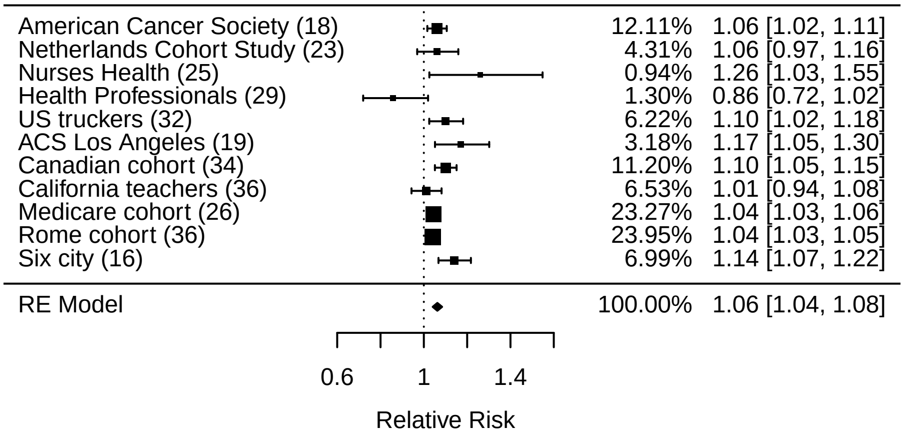
An I^2 statistic describes the percentage of variation across studies that is due to heterogeneity rather than chance. Unlike Q it does not inherently depend upon the number of studies considered. A value of 0% indicates no heterogeneity, 25% indicates low heterogeneity, 50% indicates moderate heterogeneity, and 75% indicates high heterogeneity. Values can never reach 100% and values above 90% are very rare.

Activities

- Is the pooled effect significant?
- By looking at the I^2 statistic, can you describe the level of heterogeneity?
- By looking at Cochran's Q statistic, is the heterogeneity significant?

We can create a visual summary of the meta-analysis. A forest plot is a good graphic representation of estimated results from other studies addressing the same question, along with the overall results. We can create a forest plot using object mod created above by using it as an input to `forest()` function.

```
# Creating a forest plot
forest(mod, # Model
       showweights = TRUE, # Show the contributions from each study
       transf = exp, # Transforming from log relative risk
       refline = 1, # Reference line, 1 is no increased in risk
       xlab = 'Relative Risk') # x-axis label
```

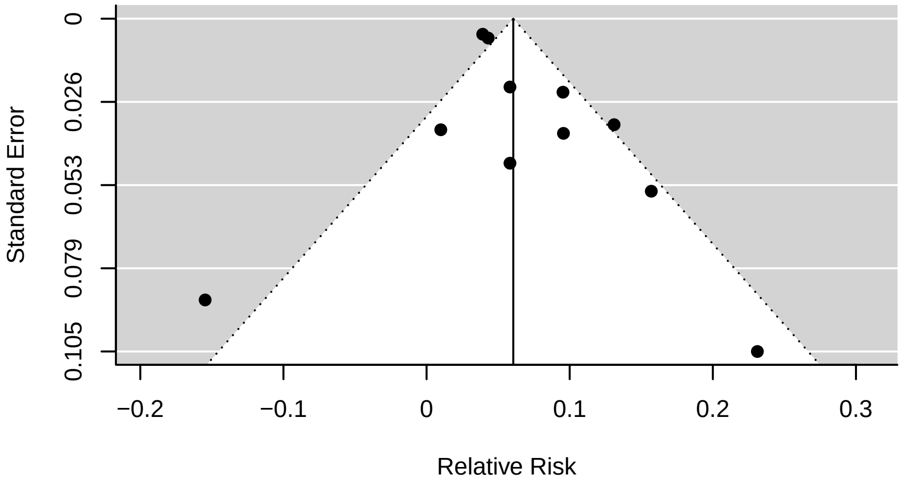


Activities

- Do any results look strange?
- Which studies have the most and least influence on the pooled effect?
- Why do you think this is? You may want to refer back to Table 1 of Hoek *et al.* (2013)

A funnel plot is designed to check the existence of publication bias. It assumes that the largest studies will be near the average, and small studies will be spread on both sides of the average. Variation from this assumption can indicate publication bias. We can create a funnel plot using object mod created above by using it as an input to `funnel()` function.

```
# Creating a funnel plot
funnel(mod, # Model
       xlab = 'Relative Risk', # x-axis label
       ylab = 'Standard Error') # y-axis label
```



Activities

- Do any results look strange?
- Is there evidence of publication bias?
- If so, study is it coming from?

Repeat this analysis but use a fixed effects model for the meta-analysis rather than a random effects model. Are there any differences in the results? If so, what? And why do you think this is?