noah_method

September 28, 2023

1 ClassAct - Noah Method

We are interested in estimating the difference in illness sessions as a proportion of attended sessions, so called "illness ratios" between schools with HEPA filters and those without. The model we have currently adopted calculates illness ratios as a function of the presence of HEPA filters - whilst we can calculate a simple difference between the predicted ratios of HEPA filtered and non-HEPA filtered schools, this does not provide us with an estimate of the variance of, and consequently a confidence interval for, this figure. We propose an alternative method to estimate the proportion of illness ratios between HEPA filtered and control schools, whereby this figure is directly estimated as a coefficient of a re-formulated poisson regression.

It should be noted that, by adding a direct estimate of the proportion of illness ratios between HEPA filtered and control schools, we remove two important features of the original model: * **Covariates:** This method does not allow for the inclusion of the covariates in our initial model, CO_2 and positive tests in the local area, and thus the estimated differences are for unadjusted illness ratios. * **HEPA/control Illness ratios:** This method decomposes the illness ratio estimates in one of HEPA/control schools into multiple parameters. As such, it is not possible to estimate the variance of the mean illness ratios for both HEPA and control schools using this model, only the variance of the difference between them. As such, this method is proposed as an additional model for estimating the overall mean proportion of illness ratios between HEPA and control schools, and not as a replacement for the original model.

1.1 The Model

We begin with a model that estimates simple count of school sessions y_i as a poisson random variable, whose rate parameter is the mean session count:

y_i ~ pois(y_bar)

The rate parameter y_bar is given by a log-linear model, with coefficients describing the type of school session (illnesss or in-person), the presence or absence of HEPA filtration, and an interaction term for session type and presence of HEPA:

log(y_bar) = b_0 + b_1*1_illness?* + *b_2*1_HEPA? + b_3*1_illness?*1_HEPA?

Indicator functions encode a simple binary for the type of session count being estimated: in-person (1_illness? = 0) or or illness (1_illness? = 1), control (1_HEPA? = 0) or HEPA (1_HEPA? = 1). The log-linear relationship between the mean session count y_bar and the coefficients, means that the log-linear model is decomposed as follows:

y_bar = e^b_0 * e^(b_1*1_illness?)* e^(b_2*1_HEPA?)* e^(b_3*1_illness?*1_HEPA?)

Table XXXX contains lay-descriptions of each of the statistics estimated by the exponents of these coefficients. We see that by taking the exponent of the b_3 coefficient, we arrive at an estimate of the the mean illness ratio in HEPA filtered schools as a proportion of the mean illness ratio in control schools. The standard methodology for estimating the variance of regression coefficients then allows us to estimate a confidence interval for this statistic.

Coefficient	Variable	Description
b_0	n/a	The mean number of in-person sessions for non-HEPA schools
b_1	1_illness?	The mean illness ratio in non-HEPA schools
b_2	1_HEPA?	The mean ratio of in-person sessions in HEPA schools to in-person sessions in control schools
b_3	1_illness?*1_hepa?	The mean proportion of illness ratios (illness sessions as a proportion of in-person sessions) in HEPA schools compared to mean illness ratios in control schools

1.2 Method

We started by re-formatting our data to facilitate this new analysis. We took each observation from the initial analysis, the mean ratio of illness sessions to in-person sessions for each school over the study period, and decomposed them into two observations: a mean count of the attended sessions and a mean count of the illness sessions. We added binary indicators for the count type (illness or in-person) and the presence or absence of HEPA filters. An example of our re-formatted data can be seen in FIGURE XXXX (see code outputs below).

We round our analyses in R, specifying our poisson log-linear model using the built-in `glm` function as follows:

```
model <- glm(
    attendance_count ~ is_illness + is_illness*hepa_filters,
    data = data,
    family = poisson(link=log)
)
```

and produced statistics using the `summary` and `confint` built-in helper functions.

2 Results

As can be seen from the below outputs, the proportion of illness ratios between HEPA and control schools is 0.769 (95% CI: 0.758, 0.780) - that is, HEPA filtered schools show an estimated reduction of sessions missed due to illness as a proportion of sessions attended of 23.1% (95% CI: 22.0%, 24.2%).

As stated above, we can provide variance estimates for the mean illness ratios of only control or HEPA schools depending on the parametarisation of this model (in this case control schools). As such, we will omit these estimates and refer the reader to the initial model for estimates of these statistics.

3 Additional Materials

3.1 Explanation by Intuition

This can be understood by intuition: the model begins at a baseline (whereby both illness_? and HEPA_? are zero), giving the mean count of in-person attendance sessions in control schools, e^b_0. We then have a coefficient to vary the count by the type of session (where illness_? = 1, HEPA_? = 0), which gives the mean proportion of illness session to in-person sessions e^b_1 in control schools.

We then have two coefficients that act in partnership to estimate the effect of HEPA filtration, b_2 and b_3. b_2 adjusts the number of sessions counted based on the presence of HEPA filters, HEPA_? = 1. In a simple model with only independent interactions between attendance type and the presence of hepa filters, the adjustment of b_2 would be unidirectional - this would imply an effect of HEPA filters as reducing/increasing **both** in person attendance and illness sessions by the same proportion. Because of this, an interaction term b_3 is included, adjusting the proportion of illness sessions in schools with HEPA filters. As such, e^b_2 is the mean proportion of in-person attended sessions between control and HEPA schools (where illness_? = 0, HEPA_? = 1) and e^b_3, only being present with both b_1 and b_2 (HEPA_? = 1, illness_? = 1), encodes the mean proportional difference in the proportion of illness to in-person sessions between control and HEPA schools.

3.2 Explanation by ?proof?

See "Noah method explainer.docx.pdf"

[1]:
```
library(bigrquery)
library(ggplot2)
library(lubridate)
library(tidyverse)
library(marginaleffects)

project_id="yhcr-prd-phm-bia-core"
attendance_sql <- "SELECT * FROM `yhcr-prd-phm-bia-core.CB_CLASS_ACT.
 ↪attendance`"
attendance_table <- bq_project_query(project_id, attendance_sql)
```

```r
start_date <- as.Date("2021-09-01")
end_date <- as.Date("2022-04-01")

hepa_school_codes <- c("H01", "H02", "H03", "H04", "H05", "H06", "H07", "H08",
                       "H09", "H10", "H11")
control_school_codes <- c("C01", "C02", "C03", "C04", "C05", "C09", "C10",
                          "C11", "C12", "C13", "C14")
study_schools <- c(hepa_school_codes, control_school_codes)

agg_data <- bq_table_download(attendance_table) %>%
    filter(School_AnonID %in% study_schools) %>%
    filter(pct_in_school > 0) %>%
    filter(Unk / (pupils * 14) < 0.01) %>%
    filter(WeekStart < end_date) %>%
    mutate(arm = case_when(School_AnonID %in% hepa_school_codes ~ "HEPA",
                           School_AnonID %in% control_school_codes ~ "Control"),
           mth = factor(month.abb[month(WeekStart)],
                        levels=c("Sep", "Oct", "Nov", "Dec", "Jan", "Feb",␣
 ↪"Mar"),
                        ordered=TRUE),
           illness_rate = prop_absent_ill * 100) %>%
    group_by(WeekStart, arm) %>%
    mutate(outlier_threshold = mean(prop_absent_ill) + 3.25 *␣
 ↪(IQR(prop_absent_ill, na.rm=TRUE)),
           is_outlier = prop_absent_ill > outlier_threshold) %>%
    filter(!is_outlier) %>%
    group_by(School_AnonID, arm) %>%
    summarise(in_school = sum(in_school),
              ill = sum(ill)) %>%
    pivot_longer(cols=c(in_school, ill),
                 names_to = "attendance_type",
                 values_to = "attendance_count") %>%
    mutate(is_illness = as.integer(attendance_type == "ill"),
           hepa_filters = as.integer(arm == "HEPA")) %>%
    ungroup()

ratio_model <- glm(
    attendance_count ~ is_illness + is_illness*hepa_filters,
    data = agg_data,
    family = poisson(link=log)
)
```

```
Attaching package: ‘lubridate’


The following objects are masked from ‘package:base’:
```

```
    date, intersect, setdiff, union


  Attaching core tidyverse packages                tidyverse
2.0.0
 dplyr   1.1.3      stringr 1.5.0
 forcats 1.0.0      tibble  3.2.1
 purrr   1.0.2      tidyr   1.3.0
 readr   2.1.4
  Conflicts
tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()
 Use the conflicted package
(<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
`summarise()` has grouped output by 'School_AnonID'. You can override
using the
`.groups` argument.
```

[2]:
```
head(agg_data, 10)
```

A tibble: 10 × 6

School_AnonID <chr>	arm <chr>	attendance_type <chr>	attendance_count <int>	is_illness <int>	hepa_filters <int>
C01	Control	in_school	87438	0	0
C01	Control	ill	4740	1	0
C02	Control	in_school	92389	0	0
C02	Control	ill	6734	1	0
C03	Control	in_school	105358	0	0
C03	Control	ill	3723	1	0
C04	Control	in_school	133913	0	0
C04	Control	ill	7584	1	0
C05	Control	in_school	163980	0	0
C05	Control	ill	6504	1	0

[4]:
```
summary(ratio_model)
```

```
Call:
glm(formula = attendance_count ~ is_illness + is_illness * hepa_filters,
    family = poisson(link = log), data = agg_data)

Coefficients:
                        Estimate Std. Error  z value Pr(>|z|)
(Intercept)           11.4192030  0.0009992 11428.30   <2e-16 ***
is_illness            -2.9588688  0.0044993  -657.62   <2e-16 ***
hepa_filters          -0.1057375  0.0014897   -70.98   <2e-16 ***
```

```
is_illness:hepa_filters -0.2623572  0.0072149   -36.36   <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: 2279616  on 41  degrees of freedom
Residual deviance:  322990  on 38  degrees of freedom
AIC: 323484

Number of Fisher Scoring iterations: 4
```

[5]:
```
illness_coef <- unname(coef(ratio_model)["is_illness"])
illness_hepa_coef <- unname(coef(ratio_model)["is_illness:hepa_filters"])
hepa_illness_ratio <- exp(illness_coef + illness_hepa_coef)
hepa_illness_ratio
```

0.039906100419378

[6]:
```
control_illness_ratio <- exp(illness_coef)
control_illness_ratio
```

0.0518775665464829

[7]:
```
exp(illness_hepa_coef)
```

0.769236166534946

[9]:
```
exp(confint(ratio_model)["is_illness:hepa_filters",])
```

```
Waiting for profiling to be done…
```

2.5 \% 0.758429637138849 **97.5 \%** 0.780185815405788

3.3 Responses to “Checking the averages”

if you calculated the the average in-school attendance session in the non-HEPA schools by hand, that it would match the output from:

[17]:
```
summary(ratio_mod)$coefficients %>%
    as.data.frame() %>%
    dplyr::select(Estimate) %>%
    filter(row.names(.) %in% c('(Intercept)')) %>%
    sum() %>%
    exp()
```

91053.5454552535

[18]:
```
agg_data %>%
    filter(!is_illness & !hepa_filters) %>%
    select(attendance_count) %>%
    unlist() %>%
    as.numeric() %>%
    mean()
```

91053.5454545455

if you calculated the the average in-school attendance session in the HEPA schools by hand, that it would match the output from:

[ ]:
```
summary(ratio_mod)$coefficients %>%
    as.data.frame() %>%
    dplyr::select(Estimate) %>%
    filter(row.names(.) %in% c('(Intercept)', 'hepa_filters')) %>%
    sum() %>%
    exp()
```

81917.3000043322

[ ]:
```
agg_data %>%
    filter(!is_illness & hepa_filters) %>%
    select(attendance_count) %>%
    unlist() %>%
    as.numeric() %>%
    mean()
```

81917.3

if you calculated the the average illness-related absences in the non-HEPA schools by hand, that it would match the output from:

[ ]:
```
summary(ratio_mod)$coefficients %>%
    as.data.frame() %>%
    dplyr::select(Estimate) %>%
    filter(row.names(.) %in% c('(Intercept)', 'is_illness')) %>%
    sum() %>%
    exp()
```

4723.63636364811

[ ]:
```
agg_data %>%
    filter(is_illness & !hepa_filters) %>%
    select(attendance_count) %>%
    unlist() %>%
    as.numeric() %>%
    mean()
```

4723.63636363636

if you calculated the the average illness-related absences in the HEPA schools by hand, that it would match the output from:

[ ]:
```
summary(ratio_mod)$coefficients %>%
    as.data.frame() %>%
    dplyr::select(Estimate) %>%
    filter(row.names(.) %in% c('(Intercept)', 'is_illness', 'hepa_filters',␣
 ↪'is_illness:hepa_filters')) %>%
    sum() %>%
    exp()
```

3269.00000005719

[ ]:
```
agg_data %>%
    filter(is_illness & hepa_filters) %>%
    select(attendance_count) %>%
    unlist() %>%
    as.numeric() %>%
    mean()
```

3269

[11]:
```
summary(ratio_mod)$coefficients %>%
    as.data.frame() %>%
    dplyr::select(Estimate) %>%
    filter(row.names(.) %in% c('is_illness')) %>%
    sum()
```

-2.9588688260603

[ ]:
```

```