The Noah method

NOTE: This model was inspired by Noah Greifer's response to a question on stats.stackoverflow.

This model calculates a mean count value - e.g., the mean count of illness-related absences in the HEPA schools - given a combination of inputs for the count-type and school variables (Table xinputvalsx). The illness ratio for HEPA schools can be calculated by dividing the mean count of illness-related absences in the HEPA schools (inputs: count type = 1; school = 1) by the mean count of in-school attendance sessions in the HEPA schools (inputs: count type = 0; school = 1). The illness ratio for non-HEPA schools can be calculated, similarly.

We are interested in the relative (multiplicative) difference between mean illness-ratio for HEPA and non-HEPA schools, i.e. the illness ratio for HEPA schools divided by the illness ratio for non-HEPA schools. A complicated approach to this would involve inputting each of the sets of input values shown in Table xinputvalsx to output each (logarithm of) mean count and calculating the risk ratio as

$$risk\ ratio = illness\ ratio_{HEPA} \div illness\ ratio_{non-HEPA}$$

$$= \frac{\bar{y}_{illness,\ HEPA}}{\bar{y}_{attendances,\ HEPA}} \div \frac{\bar{y}_{illness,\ non-HEPA}}{\bar{y}_{attendances,\ non-HEPA}}$$

$$= \frac{\bar{y}_{illness,\ HEPA}}{\bar{y}_{attendances,\ HEPA}} \times \frac{\bar{y}_{attendances,\ non-HEPA}}{\bar{y}_{illness,\ non-HEPA}}$$

While this would give us an estimate of the risk ratio, it would not provide us with an estimate of the standard error, which we need to calculate the confidence interval.

Fortunately, the specification of our model directly provides an estimate of the risk ratio and its standard error via the count type-school interaction coefficient. This is because our model specified as

$$log\left(\bar{y}_{count|count\ type,\ school}\right) = \beta_0 + \beta_1 count\ type + \beta_2 school + \beta_3 count\ type * school$$

is equivalent to

$$\bar{y}_{count|count\ type,\ school} = e^{\beta_0 + \beta_1 count\ type + \beta_2 school + \beta_3 count\ type * school}$$

where $\bar{y}$ indicates a mean, and β_* are the regression coefficients.

We can show that the exponentiated count type-school interaction coefficient – e^{β_3} – is the risk ratio, by equating the outputted mean counts to regression coefficients:

$$illness\ ratio_{non-HEPA} = \frac{\bar{y}_{illness,\ non-HEPA}}{\bar{y}_{attendances,\ non-HEPA}} = \frac{e^{\beta_0+(1)\beta_1+(0)\beta_2+(0)\beta_3}}{e^{\beta_0+(0)\beta_1+(0)\beta_2+(0)\beta_3}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

$$illness\ ratio_{HEPA} = \frac{\bar{y}_{illness,\ HEPA}}{\bar{y}_{attendances,\ HEPA}} = \frac{e^{\beta_0+(1)\beta_1+(1)\beta_2+(1)\beta_3}}{e^{\beta_0+(0)\beta_1+(1)\beta_2+(0)\beta_3}} = \frac{e^{\beta_0+\beta_1+\beta_2+\beta_3}}{e^{\beta_0+\beta_2}} = e^{\beta_1+\beta_3}$$

$$risk\ ratio = \frac{illness\ ratio_{HEPA}}{illness\ ratio_{non-HEPA}} = \frac{e^{\beta_1+\beta_3}}{e^{\beta_1}} = e^{\beta_3}$$

Thus, we have shown that the risk ratio is equivalent to e^{β_3}. We noted earlier that the model's regression coefficient is β_3 rather than e^{β_3}, and the model outputs the natural logarithm of mean counts, $\log\left(\bar{y}_*\right)$, rather than the raw mean counts used in the previous equations. So, for the same reason that we exponentiate the model outputs to obtain mean counts, we exponentiate the count type-school regression coefficient to obtain the risk ratio.

Table xinputvalsx A description of the outputs produced for a given set of input values. Outputs must be exponentiated to be in units of counts rather than log(counts).

		Independent variables (a.k.a. inputs)		
Description	**Symbol**	**Count type**	**School**	**Count type-School**
Mean count of in-school attendance sessions for HEPA schools	$\bar{y}_{attendances,\ HEPA}$	0	0	0
Mean count of in-school attendance sessions for non-HEPA schools	$\bar{y}_{attendances,\ non-HEPA}$	0	1	0
Mean count of illness-related absences for HEPA schools	$\bar{y}_{illness,\ HEPA}$	1	0	0

Mean count of illness-related absences for non-HEPA schools	$\bar{y}_{illness,\ non-HEPA}$	1	1	1

Adding covariates, v1.1

I want to compare the illness ratios for HEPA and non-HEPA schools when we include covariates to adjust for confounding. Note: without the covariates:

$$illness\ ratio_{non-HEPA} = e^{\beta_1}$$

$$illness\ ratio_{HEPA} = e^{\beta_1+\beta_3}$$

I introduce CO2 concentration and COVID rate with their regression coefficients β_4 and β_5, respectively.

The algebra shenanigans below assume that counts from non-HEPA schools are indexed by i and counts from HEPA schools are indexed by k, where $i \neq k$. It assumes that each count has a meaningful CO2 concentration that can be attributed to it, and a meaningful COVID rate that can be attributed to it.

The model structure is

$$log\left(\bar{y}_{count|count\ type,\ school}\right) = \beta_0 + \beta_1 count\ type + \beta_2 school + \beta_3 count\ type * school + \beta_4 CO2 + \beta_5 COVID$$

The illness ratios are:

$$illness\ ratio_{non-HEPA} = \frac{\bar{y}_{illness,\ non-HEPA}}{\bar{y}_{attendances,\ non-HEPA}} = \frac{e^{\beta_0+(1)\beta_1+(0)\beta_2+(0)\beta_3+(CO2_i)\beta_4+(COVID_i)\beta_5}}{e^{\beta_0+(0)\beta_1+(0)\beta_2+(0)\beta_3+(CO2_i)\beta_4+(COVID_i)\beta_5}}$$

$$= \frac{e^{\beta_0+\beta_1+(CO2_i)\beta_4+(COVID_i)\beta_5}}{e^{\beta_0+(CO2_i)\beta_4+(COVID_i)\beta_5}} = \frac{e^{\beta_1+(CO2_j)\beta_4+(COVID_j)\beta_5}}{e^{(CO2_j)\beta_4+(COVID_j)\beta_5}} = e^{\beta_1}$$

$$illness\ ratio_{HEPA} = \frac{\bar{y}_{illness,\ HEPA}}{\bar{y}_{attendances,\ HEPA}} = \frac{e^{\beta_0+(1)\beta_1+(1)\beta_2+(1)\beta_3+(CO2_k)\beta_4+(COVID_k)\beta_5}}{e^{\beta_0+(0)\beta_1+(1)\beta_2+(0)\beta_3+(CO2_k)\beta_4+(COVID_k)\beta_5}}$$

$$= \frac{e^{\beta_0+\beta_1+\beta_2+\beta_3+(CO2_k)\beta_4+(COVID_k)\beta_5}}{e^{\beta_0+\beta_2+(CO2_k)\beta_4+(COVID_k)\beta_5}} = \frac{e^{\beta_1+\beta_3+(CO2_k)\beta_4+(COVID_k)\beta_5}}{e^{(CO2_k)\beta_4+(COVID_k)\beta_5}} = e^{\beta_1+\beta_3}$$

Note, that both illness ratios are the same as when we didn't include the CO2 concentration and COVID rate. This is because all covariates are accounted for when calculating each school group's illness ratio.

While our process inherently adjusts for confounding bias with school groups (i.e. HEPA -vs- non-HEPA), collider bias can persist if assignment to HEPA or non-HEPA schools was not random.

Adding covariates, v2

I want to compare the illness ratios for HEPA and non-HEPA schools when we include covariates to adjust for confounding. Note: without the covariates:

$$illness\ ratio_{non-HEPA} = e^{\beta_1}$$

$$illness\ ratio_{HEPA} = e^{\beta_1+\beta_3}$$

I change the names of the independent variables to match SR's nomenclature:

- $count\ type = is\_illness$
- $school = hepa\_filters$

I introduce CO2 concentration and COVID rate with their regression coefficients β_4 and β_6, respectively. I introduce interaction terms for these with $is\_illness$ and $heap\_filters$, noted as regression coefficients β_5 and β_7.

The algebra shenanigans below assume that counts from non-HEPA schools are indexed by i and counts from HEPA schools are indexed by k, where $i \neq k$. It assumes that each count has a meaningful CO2 concentration that can be attributed to it, and a meaningful COVID rate that can be attributed to it.

The model structure is

$$log\left(\bar{y}_{count|count\ type,\ school}\right) = \beta_0 + \beta_1 is\_illness + \beta_2 hepa\_filters + \beta_3 is\_illness * hepa\_filters + \beta_4 co2 + \beta_5 co$$

The illness ratio for the non-HEPA schools is:

$$illness\ ratio_{non-HEPA} = \frac{\bar{y}_{illness,\ non-HEPA}}{\bar{y}_{attendances,\ non-HEPA}}$$

…written in terms of regression coefficients…

$$= \frac{e^{\beta_0+(1)\beta_1+(0)\beta_2+(0)\beta_3+(co2_i)\beta_4+(1)(co2_i)\beta_5+(covid_i)\beta_6+(1)(covid_i)\beta_7}}{e^{\beta_0+(0)\beta_1+(0)\beta_2+(0)\beta_3+(co2_i)\beta_4+(0)(co2_i)\beta_5+(covid_i)\beta_6+(0)(covid_i)\beta_7}}$$

…remove 0 terms…

$$= \frac{e^{\beta_0+\beta_1+(co2_i)\beta_4+(co2_i)\beta_5+(covid_i)\beta_6+(covid_i)\beta_7}}{e^{\beta_0+(co2_i)\beta_4+(covid_i)\beta_6}}$$

…cancel above what is below…

$$= e^{\beta_1+(co2_i)\beta_5+(covid_i)\beta_7}$$

…minor rearranging for reasons that will become apparent later…

$$= e^{\beta_1+co2_i(\beta_5)+covid_i(\beta_7)}$$

The next illness ratio is for the HEPA schools:

$$illness\ ratio_{HEPA} = \frac{\bar{y}_{illness,\ HEPA}}{\bar{y}_{attendances,\ HEPA}}$$

…written in terms of regression coefficients…

$$= \frac{e^{\beta_0+(1)\beta_1+(1)\beta_2+(1)\beta_3+(co2_k)\beta_4+(1)(co2_k)\beta_5+(covid_k)\beta_6+(1)(covid_k)\beta_7}}{e^{\beta_0+(0)\beta_1+(1)\beta_2+(0)\beta_3+(co2_k)\beta_4+(0)(co2_k)\beta_5+(covid_k)\beta_6+(0)(covid_k)\beta_7}}$$

…remove 0 terms…

$$= \frac{e^{\beta_0+\beta_1+\beta_2+\beta_3+(co2_k)\beta_4+(co2_k)\beta_5+(covid_k)\beta_6+(covid_k)\beta_7}}{e^{\beta_0+\beta_2+(co2_k)\beta_4+(covid_k)\beta_6}}$$

…extract variable names as common factors…

$$= \frac{e^{\beta_0+\beta_1+\beta_2+\beta_3+co2_k(\beta_4+\beta_5)+covid_k(\beta_6+\beta_7)}}{e^{\beta_0+\beta_2+co2_k(\beta_4)+covid_k(\beta_6)}}$$

…cancel above what is below…

$$= e^{\beta_1+\beta_3+co2_k(\beta_5)+covid_k(\beta_7)}$$

Now, we calculate the risk ratio, which is the illness ratio for HEPA schools divided by the illness rate for non-HEPA schools:

$$risk\ ratio = \frac{illness\ ratio_{HEPA}}{illness\ ratio_{non-HEPA}}$$

…written in terms of regression coefficients…

$$= \frac{e^{\beta_1+\beta_3+co2_k(\beta_5)+covid_k(\beta_7)}}{e^{\beta_1+co2_i(\beta_5)+covid_i(\beta_7)}}$$

…first pass of cancelling above what is below…

$$= \frac{e^{\beta_3+co2_k(\beta_5)+covid_k(\beta_7)}}{e^{co2_i(\beta_5)+covid_i(\beta_7)}}$$

…the variables $co2_k$, $co2_i$, $covid_k$, and $covid_i$ do not cancel because they are referring to different portions of the data set. Previously, when we replaced $is\_illness$ with 1 or 0, what we were actually doing was replacing $is\_illness$ with the mean value for that subgroup. Of course, by definition, everyone in the subgroup defined by $is\_illness = 1$ has a value of 1, so the mean value is 1.

But, things are different for CO2 concentration and COVID rate because it not everyone in the subgroups defined by $is\_illness$ and $hepa\_filters$ has the same value. Instead, for example, $co2_k$ represents the mean CO2 concentration in HEPA schools, while $co2_i$ represents the mean CO2 concentration in non-HEPA schools. Instead of cancelling $co2_k$ and $co2_i$ above and below the line, we have to incorporate them into the exponent…

$$= e^{\beta_3+co2_k(\beta_5)+covid_k(\beta_7)-(co2_i(\beta_5)+covid_i(\beta_7))}$$

$$= e^{\beta_3+co2_k(\beta_5)+covid_k(\beta_7)-co2_i(\beta_5)-covid_i(\beta_7)}$$

$$= e^{\beta_3+co2_k(\beta_5)-co2_i(\beta_5)+covid_k(\beta_7)-covid_i(\beta_7)}$$

…extract coefficients as common factors…

$$= e^{\beta_3+\beta_5(co2_k-co2_i)+\beta_7(covid_k-covid_i)}$$

Discussion

Without the covariates for CO2 concentration and COVID rate, the risk ratio is represented by the coefficient for the interaction term between $is\_illness$ and $hepa\_filters$, β_3. With the covariates included, risk ratio is represented by the sum of three components:

1. the interaction term between $is\_illness$ and $hepa\_filters$, β_3;
2. the interaction term between $is\_illness$ and $co2$, β_5, multiplied by the difference in the mean CO2 concentration observed HEPA and non-HEPA schools;
3. the interaction term between $is\_illness$ and $covid$, β_7, multiplied by the difference in the mean CO2 concentration observed HEPA and non-HEPA schools.

Calculating a point estimate for this risk ratio is easy; just sum the three components. The issue we now face is that we had an estimate of the standard error for the risk ratio in the no-covariate model, but we don't and have an estimate for the standard error for the new model that includes the covariates. This is because the risk ratio from the no-covariate model was simply the value of the regression coefficient – this coefficient had an associated standard error. In contrast, the risk ratio in the model with covariates includes a sum of coefficients each with their own standard errors (plus some extra multiplication for the fun). The algebraic gymnastics of the Noah model are now moot.

Conclusion

Simply adding covariates with main and interaction effects to the Noah model means we no longer get an estimate of the standard error for the risk ratio. To get the standard error for this new risk ratio, we would need to respecify the model.

Confidence intervals for linear combinations

Including the covariates means that we no longer have only one coefficient to represent the risk ratio. Therefore, we have to combine the standard errors of multiple coefficients to calculate the combined standard error. We should be able to do this fairly easily using these methods. I'm not 100% clear on what the estimated risk ratio would actually be, but it makes sense that it would be $\beta_3(1) + \beta_5\left(co2_k - co2_i\right) + \beta_7\left(covid_k - covid_i\right)$ because, in the unadjusted model, it was $\beta_3(1)$.