## Practical: MAIHDA analysis of diabetes

Part of the resource: <https://www.ncrm.ac.uk/resources/online/all/?id=20849>

This practical will follow the process outlined in more detail in:

Evans, Leckie, Subramanian, Bell and Merlo (2024) A tutorial for conducting intersectional multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA). *SSM – Population Health*, 26, 101664, <https://doi.org/10.1016/j.ssmph.2024.101664>

In this exercise sheet, we will go through the key steps of running a MAIHDA analysis, and understanding the outputs produced. However, we would encourage you to read the full paper, to understand a bit more depth in terms of the theoretical and methodological underpinnings of the analysis that we are doing.

The data used here are simulated, but are designed to produce realistic features of equivalent data from the United States of America.

The outcome of interest is HbA1c. This is a biomarker commonly used as an indicator of blood glucose control and diabetes. Generally, values of greater than 48 mmol/mol are indicative of a “diabetic range.

We additionally have “intersectional” identity characteristics:

* Sex (Male and Female),
* Ethnicity (White, Black, Hispanic),
* Education (Less than High School, Completed High School, Some college no degree, College degree or more)
* Income (Low, Low-middle, High-middle, and High income)
* Age (18-29, 30-44, 45-59, 60+)

The data can be found on that paper’s website, as well as in the google drive you have been provided with.

**Initial software setup, and data loading**

First change the current working directory to wherever you have saved the data for this practical, TutorialData.dta*.* You can do this using the cd command

. cd ["]*drive*:*directory*\_*name*["]

Next, we will load the data.

. use "TutorialData.dta", clear

**Generating the stratum ID**

The strata identifiers (sex, ethnicity, education, income, and age) have been chosen based on theory. We have done so attempting to balance (a) wanting a good amount of nuance in understanding of the different categories, and (b) wanting the sample size in each strata to be so small that it is difficult to analyse.

A simple way to produce the stratum ID variable is to use numerical codes that will be unique to each strata. We do this here by creating a set of 5-digit codes, where each digit corresponds to a different variable.

. generate stratum = 10000 \* sex + 1000 \* race ///

+ 100 \* education + 10 \* income + 1 \* age

This will produce a variable, stratum, where the first digit corresponds to the sex of the individual, the second digit corresponds to the strata ethnicity, etc. Since this is a categorical variable, we want to turn this into a factor variable.

We may then want to analyse the stratum data in more detail. We could tabulate it, as well as create a new variable that records the number of individuals in each stratum.

. tabulate stratum

. bysort stratum: generate strataN = \_N

**Descriptive statistics**

It would be sensible to further consider the descriptive statistics of each of our variables, to ensure they are as we would expect. For the categorical stratum variables, these would be tabulations:

. tabulate sex

. tabulate race

. tabulate education

. tabulate income

. tabulate age

Whereas for the outcome variable, we would look at continuous measures of mean and spread

. summarize HbA1c

We could additionally make a plot of our outcome variable as well.

. histogram HbA1c

**Running our two MAIHDA models**

We can now run our two MAIHDA models – first a null model, with just an intercept in the fixed part of the model and random intercepts on strata, and then a main effects model, with main effects of the strata-defining variables included. Starting with the first model:

. mixed HbA1c || stratum:, reml

We are using the mixed command, and then defining the outcome variable (HbA1c). We then enter the stratum effects as random effects by specifying stratum to be level-2 in the model.

Store the model results

. estimates store model1A

It will be useful for later to make predictions based on this model, for which we can use the predict function.

. predict m1Am, fitted

Next, let’s run the main effects model

. mixed HbA1c i.sex i.race i.education i.income i.age ///

|| stratum:, reml

Store the model results

. estimates store model1B

Again, it will be helpful to make predictions based on this model.

. predict m1Bm, fitted

For these predictions, we additionally want to save predicted confidence intervals of those predictions. This is not straightforward to do in Stata. The confidence intervals will depend on both the uncertainty associated with the regression coefficients, the variances, and the random effects and their covariates. We approximate the confidence intervals by only considering the uncertainty associated with the predicted random effects, when making plots below.

We also want to predict the stratum random effects and their standard errors

. predict m1Bu, reffects reses(m1Buse)

Finally, given that many of the differences we are looking at are between strata, it makes sense to reduce the data down to the strata level, so that we can more efficiently plot strata differences.

. collapse (mean) HbA1c, ///

by(stratum sex race education income age strataN ///

m1Am m1Bm m1Bu m1Buse)

This will take the mean of of HbA1c, and use the remaining variables as the grouping variables (these variables all vary at the strata level, so this is effectively going to group by strata).

**Model analysis**

We can now analyse our model results. A convenient way to produce a regression results table is the etable command.

. etable, estimates(model1A model1B)

Looking at model 1A, we can see that the precision-weighted stratum mean of HbA1c is 40.79. Note that this will differ slightly from the overall sample mean, since it is the stratum mean weighted by stratum size, rather than the sample mean (weighted evenly across individuals). We can also see the stratum- and individual-level variances, which are 9.33 and 90.26 respectively. On the basis of this, we can calculate the VPC

VPC = 9.33 / (9.33 + 90.26)

That is, approximately 9.4% of the variance in HbA1c can be found at the stratum level.

Model 1B introduces the additive main effects of the stratum-level variables. We can see that individuals with the highest HbA1c levels are generally Male, Black, Low Educated, relatively low income, and old. It’s worth noting the differences in effect size of these variables, though, with a very pronounced effect of being 60+ and black that are much larger than the other variables. Note, as well, that it is not always appropriate to consider these inequalities through an intersectional lens. Sometimes these differences will be produced through biological effects, whereas at other times they may be produced through the effects of social injustice and discrimination. The model cannot tell us how these effects were produced.

We can see that, comparing model 1A and 1B, the stratum-level variance reduces substantially, from 9.33 to 0.80. This is a reduction of 91.4% - this statistic is the proportional change in variance, and tells us that ~90% of the inequalities between strata is driven by additive, rather than multiplicative effects. This might seem like a lot, but it about average for what is seen in MAIHDA studies in the literature, and the remaining 8.6% may well be very important in terms of shaping intersectional inequalities.

**Looking at specific strata**

The above model outputs tell us about intersectionality *generally* – that is: how big are inequalities generally, and to what extent are those inequalities additive or multiplicative. It would now be sensible to look at specific strata- that is, what strata would we expect to have the highest levels of HbA1c, and which strata the lowest.

We can first make a plot of our strata predictions, using the stratum\_level dataframe previously created. First, we create a rank variable for our strata estimates:

. egen m1Bmrank = rank(m1Bm)

We can then plot a caterpillar plot of our main effects model’s predictions against this rank. We have previously produced these predictions and their standard errors, so this plot is fairly easy to produce:

. serrbar m1Bm m1Buse m1Bmrank, scale(1.96) yline(0)

We might want to highlight the top and bottom strata here:

. sort m1Bmrank

. list in 1/6, separator(0)

. list in -6/l, separator(0)

It can be seen that the Male, white, highly educated, high income, young group has the lowest predicted level of HbA1c, whilst the highest level of HbA1c is found in the low-income, low educated, black male group.

These results are based on both the main effects and the multiplicative differences from the main effects. However we might additionally want to consider just the multiplicative differences – that is which strata are particularly advantaged / disadvantaged in comparison to what would be expected based on just their main effects.

First we rank the predicted stratum random interaction effects from the main effects model

. egen m1Burank = rank(m1Bu)

We can then plot these using the serrbar command

. serrbar m1Bu m1Buse m1Burank, scale(1.96) yline(0)

It can be seen from this “caterpillar plot” there are a few strata which are particularly high and low, although most overlap the 0 line suggesting they are not significantly different from the mean expected by their additive effect. We could next plot only those strata which have statistically significant multiplicative effects.

To do this, we first generate a binary indicator for whether each predicted stratum random interaction effect is statistically significant.

. generate sig = (m1Bu + 1.96 \* m1Buse < 0 ///

| m1Bu - 1.96 \* m1Buse > 0)

If the expression within the parentheses is evaluated as true, then a value of 1 is assigned to the new variable sig, otherwise a value of 0 is assigned.

We then keep the subset of strata which show statistically significant predicted stratum random interaction effects

. keep if sig == 1

Then we rank these remaining predicted stratum random interaction effects

. egen m1Buranknew = rank(m1Bu)

We can then reproduce our plot

. serrbar m1Bu m1Buse m1Buranknew, scale(1.96) yline(0)

It is helpful to add a few options to show the stratum ID numbers

. serrbar m1Bu m1Buse m1Buranknew, scale(1.96) yline(0) ///

addplot((scatter m1Bu m1Buranknew, ///

mlabel(stratum) mlabposition(12) mlabgap(\*20))) ///

legend(off)

It can be seen that strata 21223 (Male, White, High-School Education, Low-middle income, age 45-59) has the biggest positive multiplicative effect, and strata 22114 (Male, Black, Low education, low income, over 60) has the biggest negative multiplicative effect. However, this is not to say that those strata are particularly advantaged or disadvantaged, just that they are more (dis)advantaged than we might expect given their combination of additive identity characteristics. As such, these results should only be interpreted in the context of the main effects in model 1B and the overall predictions produced earlier.

National Centre for Research Methods (NCRM)  
Social Sciences  
Murray Building (Bldg 58)  
University of Southampton  
Southampton SO17 1BJ  
United Kingdom

**Web** www.ncrm.ac.uk   
**Email** info@ncrm.ac.uk  
**Tel** +44 23 8059 4539  
**Twitter** @NCRMUK