eAppendix 2. Structural nested model estimation.

Below we introduce the structural nested model, all of the relevant notation, how to produce the confounding weights, and we then show the estimation algorighm that is used to produce the IV paramater estimates and 95% confidence intervals.[1] We recommend using this in conjuction with the actual SAS code presented in eAppendix 1.

The structural nested model

The structural nested model is:

$$h(E^{W1}(Y_{ij}(a)|A_i{=}a,\ R_i)) - h(E^{W1}(Y_{ij}(0)|A_i{=}a,\ R_i)) = a_v\xi$$

$Y_{ij}(a)$ represents the potential outcome for the jth pupil in the ith school at some observed adherence level a. A_i represents either a categorical or continuous school-level adherence variable. For example, in the water available schools where we have three study arms we let A_i=0 when the ith school adequately adhered to zero of the three WaSH components, let A_i=1 when the ith school adequately adhered to one or two components, and we let A_i=2 when the ith school adequately adhered to all three components. a_v represent a vector for the categorical adherence variable. R_i represents a categorical school-level randomization variable. For example, we let R_i represent a categorical variable denoting randomization to one of the three study arms in the water available group. E^{W1} represents a weighted expectation, which accounts for individual-level confounders using the weight W_{ij1}. *h* represents a link function (e.g. $h(p) = p$; $h(p) = \log(p)$; $h(p) = \log(p/(1-p))$). ξ represents a causal effect—for example a RD, logRR, or logOR corresponding to the link function that was used to transform the left parts of the model.

Producing the weights

We must first produce the overall weights (W_{ij}). W_{ij} is the product of confounding weight (W_{ij1}) and the sampling weights (W_{ij2}). W_{ij1} is the weight that is used to remove the association between individual-level confounders and randomization. W_{ij2} is the inverse of the probability of selection of each pupil into the study, and is necessary here because our study used a complex sample design. W_{ij1} and W_{ij} are produced in SAS as shown:

```
proc surveylogistic data=diarrhea;
        class pupil_grade;
        model R = pupil_grade / link=glogit;
        output out=invweights predicted=predprobs;
        weight Wij2;
        strata stratum;
        cluster psu;
run;

data weights;
        set invweights;
        Wij1 = .;
        Wij1 =1/predprobs;
        Wij = Wij1* Wij2;
        if _LEVEL_ = R;
run;
```

The structural nested model estimation algorithm

The estimating equations

To estimate the parameters of interest in the structural nested mean model, we used an iterative algorithm which applies Newton’s method. The two estimating equations are shown below:

(Equation 1) $\sum_i \sum_j \mathrm{W}_{ij}\, \mathrm{D}_i^{\mathrm{T}} \left[\mathrm{Y}_{ij} - \mu\,(\mathrm{A}_i,\ \mathrm{R}_i;\ \eta)\right] = 0$

(Equation 2) $\sum_i \sum_j \mathrm{W}_{ij}\, \mathrm{R}_{vi}^{\mathrm{T}}\ [\mathrm{h} - 1\ (h(\mu(\mathrm{A}_i, \mathrm{R}_i;\ \eta)) - A_{vi}\xi) - \alpha]\ =\ 0$

η represents the $E^{W1}(Y_{ij}\ |A_{i,}\ R_i)$, α represents the $E^{W1}(Y_{ij}(0))$, ξ represents the causal effect of adherence on the outcome given A_i and R_i, and D is a function of A_i and R_i, defined as $(A_{vi}, R_{vi}, A_i{*}R_i)^T$, R_{iv} is a vector of dummy variables representing randomization to one of three arms, and all other variables are as previously defined. A SAS program which was designed for a three armed trial with two strata was obtained from Brumback,[1] and was slightly modified to allow for variation in either the number of strata or the number of study arms. Generally, the steps to solving these estimating equations are shown below.

Solving the estimating equations

Step 1. We first solve the estimating equation 1 using a fully parameterized model, to obtain an estimate of η for each participant. For instance, if our outcome followed a binomial distribution, we might use PROC GENMOD as shown:

```
proc genmod data=sim.data0;
        model Y= A1 A2 R1 R2 A1*R1 A1*R2 A2*R1 A2*R2/dist= bin link=logit;
        weight Wij;
        output out=sim.xbeta xbeta=linpred;
run;
```

Step 2. Letting $h^{-1}(.) = g(.)$, substitute $\hat{\eta}$ from equation 1, for η in equation 2. Using Newton’s method, we linearize $g(D_i\, \hat{\eta} - A_{vi}\xi)$ about an initial (or current) estimate of ξ, ξ^t, where t indexes the iteration number. Equation 2 reduces to:

(Equation 2a) $\sum_i \sum_j \mathrm{W}_{ij}\, \mathrm{R}_{vi}^{\mathrm{T}}\ [\mathrm{g}(\mathrm{D}_i\ \hat{\eta} - A_{vi}\xi) - \alpha]\ =\ 0$

$g(D_i\, \hat{\eta} - A_i\xi)$ is approximated by $(Y_i{*} - A_{vi}{*}\xi^t)$, where $Y_i{*}$ and $A_{vi}{*}$ are derived using Taylor series approximation. For instance, for the logistic structural nested model, g(x) ≡

exp(x)/(1+exp(x)), and we let $Y_i^* \equiv g(D_i \hat{\eta} – A_{vi}\xi^t) + A_{vi}^*\xi^t$, and $A_{vi}^* \equiv A_{vi} g(D_i \hat{\eta} – A_{vi}\xi^t)(1– g(D_i \hat{\eta} – A_{vi}\xi^t))$. Equation (2) further reduces to:

(Equation 2b) $\sum_i \sum_j \mathrm{W}_{ij}\ \mathrm{R}_{vi}^{\mathrm{T}}\ [\mathrm{Y}_i^* - \mathrm{A}_{vi}^* \xi - \alpha] = 0$

For example, when using the logistic structural nested model Y_i^* and A_{vi}^* can be calculated within a data step in SAS using the linear predictor (output in step 1), the adherence variables (A_{vi}), the outcome variable (Yi), and an initial estimate of the causal effect (ξ^t) using the code:

```
lp=linpred-A1*squig1 - A2*squig2;
expitlp=exp(lp)/(1+exp(lp));
Ystar = expitlp + (A1*squig1 + A2*squig2)*expitlp*(1-(expitlp));
Astar1 = (A1*expitlp)*(1-(expitlp));
Astar2 = (A2*expitlp)*(1-(expitlp));
```

If we were instead using the log structural nested model, we would let g(x) ≡ exp(x), and we let $Y_i^* \equiv g(D_i \hat{\eta} – A_{vi}\xi^t) + A_{vi}^*\xi^t$, and $A_{vi}^* \equiv A_{vi} g(D_i \hat{\eta} – A_{vi}\xi^t)$. Y_i^* and A_{vi}^* would then be calculated using the following code within a SAS data step:

```
lp=linpred-A1*squig1 - A2*squig2;
explp=exp(lp);
Ystar = explp*(1+ A1*squig1 + A2*squig2);
Astar1 = A1*explp;
Astar2 = A2*explp;
```

Step 3. Solve equation 2b using IV software using Y_i^* as the response variable, R_i as the instrument, and A_i^* as the endogenous regressor, and obtain updated estimates of ξ^t. For example, using SAS's PROC SYSLIN:

```
proc syslin data=sim.iv 2sls;
        endogenous Astar1 Astar2;
        instruments R1 R2;
        model Ystar =Astar1 Astar2;
        weight Wij;
run;
```

Step 4. Update the initial estimate of ξ^t.

Step 5. Repeat steps 2 to 4 iteratively, until all parameters converge on a fixed value of ξ.

Calculating the IV parameter of interest

Step 6. Researchers often only target the parameter ξ, which is either a logOR, a logRR (or log prevalence ratio), or a RD, each of which is conditional on both A and R. Of particular interest, may be to calculate a prevalence ratio that is conditional only on A. Specifically, the prevalence ratio comparing the prevalence of disease among adherers to what the prevalence of disease would have been had this same group not adhered: PR(a)= ($E^{W1}(Y_{ij}(a)|A_i=a)$) / $E^{W1}(Y_{ij}(0)|A_i=a)$). The numerator of interest from this prevalence ratio, i.e. $E^{W1}(Y_{ij}(a)|A_i=a)$, is easily calculated without the structural nested model. This can be done, for example, by regressing Y_{ij} on A_i in PROC REG (while using the W_{ij} weight) and outputting the 'parameter estimates' for each participant. The denominator, $E^{W1}(Y_{ij}(0)|A_i=a)$, is a counterfactual and is calculated from the structural nested model parameters. Rather than the causal parameter (i.e. ξ), of particular interest from the structural nested model is h($E^{W1}(Y_{ij}(0)|A_i=a, R_{i,})$, which differs from our IV denominator of interest in that it is also conditional on R and also that a link function is applied to it. We can apply the inverse link function (e.g. the expit, or exponential function) to h($E^{W1}(Y_{ij}(0)|A_i=a, R_i)$ to produce a counterfactual prevalence of disease for each participant in the study. To then make these conditional prevalences marginal on only A, we regress $E^{W1}(Y_{ij}(0)|A_i, R_i)$ on A_i, again using PROC REG (while using the W_{ij} weight) and outputting the parameter estimates. The resulting prevalences, $E^{W1}(Y_{ij}(0)|A_i=a)$, represent the true potential outcome had a participant's cluster counterfactually not adhered to the intervention (e.g. had it been assigned to R=0). The IV parameter of interest (e.g. a prevalence ratio) is this numerator divided by this denominator.

Variance estimation

Step 7. To estimate the variance, we use the jackknife estimator of the variance. This is a method where we systematically delete each primary sampling unit (school) and estimate the parameter of interest without that individual school, following steps 1-6 above repeatedly for all schools. The variance is then estimated by measuring the sum of the squared differences of each estimate from the overall parameter estimate, which is multiplied by a correction factor that accounts for the stratification. The jackknife estimator is:

(Equation 3) $\hat{\text{var}}(\hat{\theta}) = \sum_{h=1}^{H}((C_h - 1)/C_h) \sum_{c=1}^{C_h}(\hat{\theta}^{hc} - \hat{\theta})^2$,

where $\hat{\theta}$ represents the overall parameter estimate and $\hat{\theta}^{hc}$ represents the paramater estimate deleting the c^th^ school which is in the h^th^ stratum (district).

REFERENCES

1. Brumback BA, He Z, Prasad M, Freeman MC, Rheingans R. Using structural-nested models to estimate the effect of cluster-level adherence on individual-level outcomes with a three-armed cluster-randomized trial. *Statistics in medicine.* Apr 30 2014;33(9):1490-1502.