Supplementary Information Part A:

'Estimating Mode Effects from a Sequential Mixed-Mode Experiment Using Structural Moment Models'

S1 Plausible Data Generating Processes and No Effect Modification

Let $\mathbf{Y}^*$ represent the latent true values of the survey variables in the analysis, $\mathbf{Y}_0, \mathbf{Y}_1$ the mode-specific potential outcomes, and $\mathbf{Y} = (1-D)\mathbf{Y}_0 + D\mathbf{Y}_1$ the observed data. We now consider two plausible data generating mechanisms for the data in a mixed-mode sequential design. Both induce non-random non-ignorable selection.

(i) Mode effects determined prior to individual mode choice.

The mode-specific potential outcomes (and hence the mode effect) depend on the latent-but-true characteristics of the individual $\mathbf{Y}^*$ and the design and implementation of the survey. Hence, $\boldsymbol{\mathcal{M}} = \mathbf{Y}_1 - \mathbf{Y}_0$ is determined after $\mathbf{Y}^*$ and M but before the individual chooses whether to comply. This data generating process has the form

$$P(\mathbf{Y}^*)\zeta(M)\zeta(\mathbf{Y}_0, \mathbf{Y}_1|\mathbf{Y}^*)\zeta(D|M, \mathbf{Y}^*, \mathbf{Y}_0, \mathbf{Y}_1). \quad \text{(S1.1)}$$

The density $P(\mathbf{Y}^*)$ is for the survey characteristics of the target population and exists independently of the study, while the densities indicated by ζ depend on the study (its design and timing) as well as the characteristics of the subjects. There is no density for $\mathbf{Y}$ because it is uniquely determined given D and $\mathbf{Y}_0$ and $\mathbf{Y}_1$ by $(1-D)\mathbf{Y}_0 + D\mathbf{Y}_1$.

The randomization density $\zeta(M)$ is trivial but the *measurement* process $\zeta(\mathbf{Y}_0, \mathbf{Y}_1|\mathbf{Y}^*)$ and the *compliance/selection* process $\zeta(D|M, \mathbf{Y}^*, \mathbf{Y}_0, \mathbf{Y}_1)$ are not.

The measurement model does not depend on M and so incorporates the usual assumption that potential mode outcomes depend on individuals' survey characteristics but not on their randomization outcomes.

However, the model for compliance is completely unconstrained: if M is a valid instrumental variable then it must be associated with the compliance decision, and the dependence on $\mathbf{Y}_0$ and $\mathbf{Y}_1$ allows compliance to depend explicitly on simple mode effect $\boldsymbol{\mathcal{M}}$ as well as an individual's survey characteristics.

Under (S1.1), the resulting mode-effect distribution among those who choose web with randomization outcome M is

$$p(\mathbf{Y}_0, \mathbf{Y}_1|D=1, M) = \int_{\mathbf{Y}^*} \frac{P(\mathbf{Y}^*)\zeta(\mathbf{Y}_0, \mathbf{Y}_1|\mathbf{Y}^*)\zeta(D=1|M, \mathbf{Y}^*, \mathbf{Y}_0, \mathbf{Y}_1)}{p(D=1|M)} \partial\mathbf{Y}^*, \quad \text{(S1.2)}$$

which will generally depend on M so that the NEM assumption does not hold.

There are two exceptions to this. *Exception 1* is when mode selection is independent of the individual's characteristics or mode outcomes such that $\zeta(D=1|M, \mathbf{Y}^*, \mathbf{Y}_0, \mathbf{Y}_1) = \zeta(D=1|M)$, in which case $p(\mathbf{Y}_0, \mathbf{Y}_1|D=1, M) = p(\mathbf{Y}_0, \mathbf{Y}_1)$.

Exception 2 is when a) the measurement model does not depend on the survey characteristics such that $\zeta(\mathbf{Y}_0, \mathbf{Y}_1|\mathbf{Y}^*) = \zeta(\mathbf{Y}_0, \mathbf{Y}_1)$, and b) compliance does not depend on the potential mode

outcomes such that $\zeta(D=1|M,\mathbf{Y}^*,\mathbf{Y}_0,\mathbf{Y}_1)=\zeta(D=1|M,\mathbf{Y}^*)$. If a) and b) hold then $p(\mathbf{Y}_0,\mathbf{Y}_1|D=1,M)=p(\mathbf{Y}_0,\mathbf{Y}_1)$ as above. While we argue that it is not implausible to assume that compliance depends only on an individual's true characteristics, it is generally unreasonable to assume that mode effects are independent of the true characteristics.

(ii) Mode effects determined immediately after individual mode choice.

Alternatively, $\mathbf{Y}_0,\mathbf{Y}_1$ are determined only *after* the participants have chosen to comply or noncomply with their randomized allocations in the data generating process, that is,

$$P(\mathbf{Y}^*)\zeta(M)\zeta(D|M,\mathbf{Y}^*)\zeta(\mathbf{Y}_0,\mathbf{Y}_1|\mathbf{Y}^*,D). \quad \text{(S1.3)}$$

Now compliance depends only on an individual's true survey characteristics and randomization, and the measurement model can depend on whether individuals comply or not with their randomization.

Under (S1.3),

$$p(\mathbf{Y}_0,\mathbf{Y}_1|D=1,M)=\int_{\mathbf{Y}^*}\frac{P(\mathbf{Y}^*)\zeta(D=1|M,\mathbf{Y}^*)\zeta(\mathbf{Y}_0,\mathbf{Y}_1|\mathbf{Y}^*,D=1)}{p(D=1|M)}\partial\mathbf{Y}^*, \quad \text{(S1.4)}$$

will generally depend on M so, again, NEM does not generally hold.

NEM again holds under trivial exception 1 above, but also under a variation on exception 2: if the measurement model does not depend on the true characteristics among those who choose web, that is, $\zeta(\mathbf{Y}_0,\mathbf{Y}_1|\mathbf{Y}^*,D=1)=\zeta(\mathbf{Y}_0,\mathbf{Y}_1|D=1)$, then $p(\mathbf{Y}_0,\mathbf{Y}_1|D=1,M)=p(\mathbf{Y}_0,\mathbf{Y}_1|D=1)$.

Finally, we compare data generating processes (i) and (ii) for the mixed-modes experiment with that underpinning classical treatment-effect problems, which have the form

$$P(\boldsymbol{\varepsilon})\zeta(Z)\zeta(D|\boldsymbol{\varepsilon},Z)\zeta(\mathbf{Y}_0,\mathbf{Y}_1|\boldsymbol{\varepsilon},D), \quad \text{(S1.5)}$$

where Z is the instrumental variable, $\mathbf{Y}_0$ and $\mathbf{Y}_1$ are respectively the potential outcomes under control and treatment , and $\boldsymbol{\varepsilon}$ represents the effect of unobserved confounding. Clearly, the unobserved confounding take the place of the true characteristics in the mode-effect data generating process (ii). Exception 2 is explicitly built-in to the classical linear casual model

$$Y_d=\beta_0+\beta_1 d+\varrho_0+(\varrho_1-\varrho_0)d,$$

where $E(\varrho_0|D=0)\neq E(\varrho_0|D=1)$ but the treatment effect

$$Y_1-Y_0=\beta_1+\varrho_1-\varrho_0$$

depends on a heterogeneity term $\varrho_1-\varrho_0$ that is explicitly taken to satisfy either $E(\varrho_1-\varrho_0|D,\boldsymbol{\varepsilon})=0$ or $E(\varrho_1-\varrho_0|D=1,\boldsymbol{\varepsilon})=E(\varrho_1-\varrho_0|D=1)$. In other words, the treatment-effect heterogeneity does not depend on unobserved confounding $\boldsymbol{\varepsilon}$ and NEM holds. This assumption accepted in the treatment-effect estimation literature but, as outlined above, it cannot be so easily justified for mixed-mode designs.

S2 Two Stage Least Squares and G-estimation

The potential outcomes Y_{dm} are taken to exist prior to the experiment taking place, the potential outcome. This represents the measurement we would obtain had the individual lived in a household that was allocated to mode m and chosen mode d. There are thus four potential

outcomes of which we observe only one: $Y = Y_{DM}$. The single unit treatment value assumption (SUTVA) is taken to hold such that Y_{dm} is taken to be independent of d and m for every other sample member (including those living in the same household): in other words, the randomization and mode choice of the other sample members has no impact on the potential outcome of each individual (Angrist et al. 1996).

For M to be a valid instrumental variable, it must satisfy the following conditions:

1. Exclusion restriction $Y_{dm} = Y_d$
2. Independence of randomization and mode-specific potential outcomes: $M \perp\!\!\!\perp (Y_0, Y_1)$
3. A non-null association exists between M and D.

Consider IV regression for the linear model

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i, \quad \text{(S2.1)}$$

where error ϵ_i represents the combined effect of every cause of Y_i not explicitly included in the linear model, and τ_1 is the causal effect of mode, that is, the effect of changing D_i from 0 to 1 while simultaneously holding everything else (i.e. ϵ_i) fixed (*ceteris paribus*). Ordinary least squares is unbiased and consistent for β_1 only if $E(\epsilon_i|D_i) = 0$, that is, D_i is 'exogenous'. However, if $E(\epsilon_i|D_i) \neq 0$ then D_i is 'endogenous' which, in this case, is due to non-random selection in which participants' mode choices are potentially related to their mode effects.

If D_i is 'endogenous' , two-stage least squares (2SLS) should be used instead of ordinary least squares. 2SLS estimation proceeds as follows: 1. Regress D_i on M_i to obtain the predicted value $\widehat{D}_i = \hat{\pi}(M_i)$, where $\pi(M_i) = \Pr(D_i = 1|M_i)$; and 2. Regress of Y_i on $\widehat{D}_i$ to obtain $\hat{\beta}_1$. The 2SLS estimator in this simple case reduces to

$$\hat{\beta}_1 = \frac{E(Y_i|M_i = 1) - E(Y_i|M_i = 0)}{\pi(1) - \pi(0)}, \quad \text{(S2.2)}$$

which is consistent, but not unbiased, for β_1.

Our framework is based on the structural mean models (SMMs) for causal inference (Hernán and Robins 2006). A linear SMM for the effect of mode on the mean of survey variable $Y$is

$$E(Y_i - Y_{0i}|D_i, M_i) = \mu_1 D_i, \quad \text{(S2.3)}$$

where $\mu_1 = E(Y_{1i} - Y_{0i}|D_i = 1)$ is the average difference between the web and face-to-face mode responses among those who choose web. In contrast to IV regression, the SMM does not constrain heterogeneity in In addition to core conditions 1-3 above, the SMM model makes clear that we must also assume the mode effect among those choosing web is the same for compliers as it is for non-compliers.

The mode effect is identified by the conditional mean independence (CMI) assumption

$$E(Y_{0i}|M_i) = E(Y_{0i}), \quad \text{(S2.4)}$$

which follows under core conditions 1-2. Under SMM (S2.3), this can be rewritten as

$$E(Y_{0i}|M_i) = E\{E(Y_i|D_i, M_i) - \mu_1 D_i|M_i\} = E(Y_i - \mu_1 D_i|M_i) = E(Y_{0i}). \quad \text{(S2.5)}$$

The general form of the g-estimator for μ_1 is

$$\sum_i a_0(M_i) U_i = 0, \quad \text{(S2.6)}$$

where $U_i = Y_i - \mu_1 D_i$ and the choice of $a_0(M_i)$ must satisfy $E\{a_0(M_i)\} = 0$. Note that U_i is local notation and does **not** correspond to U_i in Submitted Paper (2020, equations (5-6)). Irrespective of this choice, the solution to (S2.6) is simply the 2SLS estimator (S2.2). Standard results for g-estimators give

$$a_0(M_i) = (\pi_i - \pi)\sigma_U^{-2}, \quad \text{(S2.7)}$$

where $\pi = \Pr(D_i = 1)$, $\pi_i = \pi(M_i)$ and $\sigma_U^2 = E(U_i)$. The asymptotic distribution is

$$\sqrt{n}(\hat{\mu}_1 - \mu_1) \sim N\{0, \sigma_U^2/\text{var}(\pi_i)\}. \quad \text{(S2.8)}$$

While this estimator would be semi-parametrically efficient if $E(U_i^2|M_i) = E(U_i^2)$ and no modelling assumptions were made about $E(Y_{0i})$, basing variance estimation on $\sigma_U^2/\text{var}(\pi_i)$ will lead to over-estimated standard errors because $E(Y_{0i})$ from (S2.5) can be included in the model without imposing further constraints on the observed data law simply by adding $\mu_0 = E(Y_{0i})$ as a parameter. Hence, the estimating equation for μ_0 and μ_1 is

$$\sum_i \mathbf{a}_0(M_i)\overline{U}_i = \mathbf{0}, \quad \text{(S2.9)}$$

where the mean of $\mathbf{a}_0(M_i)$ can be non-zero but

$$\overline{U}_i = Y_i - \mu_0 - \mu_1 D_i \quad \text{(S2.10)}$$

satisfies $E(\overline{U}_i|M_i) = 0$ under CMI (Clarke et al. 2015). Note that $\overline{U}_i$ corresponds to U_i in Submitted Paper (2020, equations (5-6); note also that (S2.10) has the same form as the residual of the 2SLS estimator (S2.1). Semi-parametric theory for (S2.9) reveals the efficient choice to be

$$\mathbf{a}_0(M_i) = \begin{pmatrix} 1 \\ \pi_i \end{pmatrix} \sigma_{\overline{U}}^{-2},$$

with asymptotic marginal distribution

$$\sqrt{n}(\hat{\mu}_1 - \mu_1) \sim N\{0, \sigma_{\overline{U}}^2/\text{var}(\pi_i)\}. \quad \text{(S2.11)}$$

The asymptotic variance above is smaller than in (S2.8) because $\sigma_{\overline{U}}^2 = E(\overline{U}_i^2) = \sigma_U^2 - \mu_0^2 \leq \sigma_U^2$.

Generally, the use of a mean-centred $\overline{U}_i$ (e.g. when covariates are included in the SMM) requires further modelling assumptions, with bias introduced if the model for $E(\overline{U}_i|M_i, \mathbf{C}_i)$ is incorrectly specified. However, the mean-centring is trivial in this case and so preferable to standard g-estimation, and will be used to construct estimators of the structural moment, variance and covariance models.

S3 Structural Moment Models

S3.1 Choice of efficient instrument for SMoMs

Denote the SMoM parameters by $\boldsymbol{\theta}$ (this includes the mean-centring parameter) and its residual ε_i satisfying $E(\varepsilon_i|M_i) = 0$. We wish to determine an 'efficient instrument' $\mathbf{a}_0(M_i)$ satisfying unconditional moment restriction

$$E\{\mathbf{a}_0(M_i)\varepsilon_i\} = \mathbf{0}, \quad \text{(S3.1)}$$

where, as in Section S2 above, $\varepsilon_i = U_i$ for a classical g-estimator or $\varepsilon_i = \overline{U}_i$ for a mean-centred GMM estimator. The efficient instrument, should it exist, is semi-parametrically efficient in that it achieves the lowest possible variance among consistent and asymptotically normal estimators where the data are constrained by the SMoM and the CMI assumptions alone.

Bowden and Vansteelandt (2011) use semi-parametric theory (e.g. Tsiatis 2006) to determine $\mathbf{a}_0(M_i)$ for linear and log-linear SMMs for non-mean-centred residuals, that is, those satisfying $E(U_i|M_i) = E(U_i) \neq 0$. Simply replacing the outcome Y with Y^k in their results (or with $X^j Y^k$), the efficient choice $\mathbf{a}_0(M_i)$ satisfying $E\{\mathbf{a}_0(M_i)\} = \mathbf{0}$ is

$$\mathbf{a}_0(M_i) = \sigma_U^{-2}(M_i)\left[E\left(-\frac{\partial U_i}{\partial \boldsymbol{\theta}}\middle| M_i\right) - E\{\sigma_U^{-2}(M_i)\}^{-1}E\left\{\sigma_U^{-2}(M_i)E\left(-\frac{\partial U_i}{\partial \boldsymbol{\theta}}\middle| M_i\right)\right\}\right],$$

where $\sigma_U^2(M_i) = E\left(U_i^2\middle|M_i\right) = E\left(U_i^2\middle|M_i\right)$ because M_i is randomised.

For mean-centred estimators, we adapt their derivation to replace the $E\{\mathbf{a}_0(M_i)\} = \mathbf{0}$ constraint with $E\{\mathbf{a}_0(M_i)\} \neq \mathbf{0}$ and $E(\overline{U}_i|M_i) = 0$ to show that

$$\mathbf{a}_0(M_i) = \sigma_{\overline{U}}^{-2}(M_i)E\left(-\frac{\partial \varepsilon_i}{\partial \boldsymbol{\theta}}\middle| M_i\right), \quad \text{(S3.2)}$$

where randomisation again leads to $\sigma_{\overline{U}}^2(M_i) = E\left(\overline{U}_i^2\middle|M_i\right) = E\left(\overline{U}_i^2\right)$, which can now be interpreted as a residual variance.

In both cases, the asymptotic distribution of the g-estimator is

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \sim N\left[\mathbf{0}, E\left\{\mathbf{a}_0(M_i)\mathbf{a}_0^T(M_i)\varepsilon_i^2\right\}^{-1}\right], \quad \text{(S3.3)}$$

where $\boldsymbol{\theta}_0$ is the true value of the SMoM.

In the mixed-mode example, because D_i is binary, the mean-centred residual is linear for both linear and log-linear SMoMs (in the latter case, $Y^k \exp(-\lambda D) = Y^k - \{1 - \exp(-\lambda)\}D$). The resulting linear model means that the estimator is the linear projection of D onto the space spanned by $\mathbf{a}_0(M_i)$, which is identical to the linear projection onto the space spanned by $(1, M_i)^T$ because M_i is binary and the efficient instrument is always a linear combination of M_i. Thus, the GMM estimator with IV M_i as the IV is semi-parametrically efficient.

The form of the efficient instrument for SVMs and SCMs is far more complex but is discussed, and the efficient instrument for the SVM set out, in Appendix SA1 below. Ultimately, the same argument about efficiency for SMoMs also follows because both M_i and D_i are binary (Clarke et al. 2015).

S3.2 Mode effect on the distribution of a nominal categorical variable

Suppose now that Y_i is a nominal categorical variable with $k+1$ categories. For such variables, it is inappropriate to think of mode effects on the mean, variance or any other mean-centred moment of its distribution. Any change to the probability of being in one category affects the entire distribution and so corresponds to a component of the overall mode effect.

We can implement this straightforwardly using the same approach as above. One begins by arbitrarily relabelling the categories of Y_i as $0, \dots, L$ with category 0 as the baseline, or reference, category against which the others are to be compared. One can then define dummy variables for each of the remaining k categories of Y_i as

$$\mathbf{Y}_i = \left(Y_i^{[1]}, \cdots, Y_i^{[L]}\right)^T,$$

where $Y_i^{[j]} = 1$ if $Y_i = j$ and $Y_i^{[j]} = 0$ if $Y_i \neq j$ for $j = 1, \dots, L$.

The effect of mode on Y_i can thus be captured by the following multivariate linear SMM:

$$E(\mathbf{Y}_i - \mathbf{Y}_{0i} | D_i, M_i) = \boldsymbol{\mu}_1 D_i, \quad \text{(S3.4)}$$

where $\boldsymbol{\mu}_1 = \left(\mu_1^{[1]}, \cdots, \mu_1^{[L]}\right)^T$ and $\mu_1^{[j]} = E\left(Y_{1i}^{[j]} - Y_{0i}^{[j]} \middle| D_i = 1\right)$ indicates the effect of mode on category j. A mode effect is thus present if there is evidence to reject the null hypothesis that

$$H_0: \mu_1^{[1]} = \cdots = \mu_1^{[L]} = 0.$$

In terms of estimation, the mean-centred residual for (S3.4) can be written

$$\overline{\mathbf{U}}_i = \mathbf{Y}_i - X_i \boldsymbol{\mu}, \quad \text{(S3.5)}$$

where $X_i = I_L \otimes (1, D_i)$ is the $L \times 2L$ design matrix, $\otimes$ is the Kronecker product, I_L is the $L \times L$ identity matrix, $\boldsymbol{\mu} = \left(\mu_0^{[1]}, \mu_1^{[1]}, \dots, \mu_0^{[L]}, \mu_1^{[L]}\right)^T$ and $\mu_0^{[j]} = E\left(Y_{0i}^{[j]}\right)$. Tsiatis (2006, Sec. 4.5) shows that the efficient score is

$$\mathbf{s}_n(\boldsymbol{\mu}) = \sum_{i=1}^{n} A_0(M_i) \overline{\mathbf{U}}_i = \sum\nolimits_i \hat{X}_i C^{-1} \overline{\mathbf{U}}_i,$$

where $\hat{X}_i = I_L \otimes \{1, \pi(M_i)\}$, $C = E(\overline{\mathbf{U}}_i \overline{\mathbf{U}}_i')$ and $\text{cov}\{\sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\} = \left\{E\left(\hat{X}_i C^{-1} \hat{X}_i^T\right)\right\}^{-1}$. SMM (S3.4) can alternatively be estimated using generalized estimating equations (Liang and Zeger 1986).

S3.3 Mode effect on the variance of a continuous variable

The effect of mode on the variance can modelled specified using the following log-linear structural variance model (SVM):

$$\log\{\text{var}(Y_i | D_i, M_i)\} - \log\{\text{var}(Y_{0i} | D_i, M_i)\} = \lambda_1 D_i, \quad \text{(S3.6)}$$

subject to CMI (S2.4) and

$$E\left(Y_{0i}^2 | M_i\right) = E\left(Y_{0i}^2\right), \quad \text{(S3.7)}$$

both of which hold under core conditions 1-2. (Note that, if Y is binary, the extra moment restriction (S3.7) adds no further information because $Y_{0i}^2 = Y_{0i}$.)

The target parameter

$$\exp(\lambda_1) = \text{var}(Y_{1i}|D_i = 1)/\text{var}(Y_{0i}|D_i = 1), \quad \text{(S3.7)}$$

is the ratio of the variances of the web responses to that of the counterfactual face-to-face responses among those who choose web. The estimator of λ_1 is (see Appendix SA1)

$$\hat{\lambda}_1 = \log\left(\frac{A}{\mu_1^2 B + 2\mu_1 C + D}\right), \quad \text{(S3.8)}$$

where μ_1 is the effect of mode on the mean from SMM (S2.3),

$$A = \pi(1)\text{var}(Y_i|D_i = 1, M_i = 1) - \pi(0)\text{var}(Y_i|D_i = 1, M_i = 0),$$

$$B = \pi(0) - \pi(1),$$

$$C = \pi(1)E(Y_i|D_i = 1, M_i = 1) - \pi(0)E(Y_i|D_i = 1, M_i = 0),$$

and

$$\begin{aligned} D = \{1 - \pi(0)\}\{&\text{var}(Y_i|D_i = 0, M_i = 0) + E^2(Y_i|D_i = 0, M_i = 0)\} \\ &- \{1 - \pi(1)\}\{\text{var}(Y_i|D_i = 0, M_i = 1) + E^2(Y_i|D_i = 0, M_i = 1)\} \\ &+ \pi(0)E^2(Y_i|D_i = 1, M_i = 0) - \pi(1)E^2(Y_i|d_i = 1, M_i = 1). \end{aligned}$$

Estimation is more straightforward using GMM based on the residual

$$V_i = \exp(-D_i\lambda_1)\epsilon_i^2 + (U_i - \epsilon_i)^2,$$

where $U_i = Y_i - D_i\mu_1$ is the residual for SMM (S2.3) and $\epsilon_i = Y_i - \beta_0 - \beta_1 M_i - \beta_2 D_i - \beta_{12} M_i D_i$ is the residual of the association model, that is, the saturated linear regression of Y_i on M_i, D_i and $M_i D_i$. This residual comes from expanding the CMI condition

$$\begin{aligned} E\left(Y_{0i}^2|M_i\right) &= E\left\{E\left(Y_{0i}^2|D_i, M_i\right)\middle|M_i\right\} = E\{\text{var}(Y_{0i}|D_i, M_i) + E^2(Y_{0i}|D_i, M_i)|M_i\} \\ &= E_{D_i}[\exp(-D_i\lambda_1)\text{var}(Y_i|D_i, M_i) + \{E(Y_i|D_i, M_i) - D_i\mu_1\}^2|M_i] \\ &= E_{Y_i, D_i}\left\{\exp(-D_i\lambda_1)\epsilon_i^2 + (U_i - \epsilon_i)^2|M_i\right\}, \end{aligned}$$

because $\text{var}(Y_i|D_i, M_i) = E\left(\epsilon_i^2\middle|D_i, M_i\right)$ and $U_i - \epsilon_i = \beta_0 + \beta_1 M_i + (\beta_2 - \mu_1)D_i + \beta_{12} M_i D_i$.

The residual V_i is analogous to U_i because it satisfies $E(V_i|M_i) = E(V_i) \neq 0$ but, as already discussed above, it is desirable to work with a zero-mean residual to efficiently estimate the standard error of $\hat{\lambda}_1$. We choose the mean-centred residual

$$V_i = \exp(-D_i\lambda_1)\epsilon_i^2 + (U_i + \mu_{10} - \epsilon_i)^2 - \mu_{20}, \quad \text{(S3.9)}$$

where $U_i = Y_i - \mu_{10} - D_i\mu_1$, $\mu_{10} = E(Y_{0i})$ and $\mu_{20} = E\left(Y_{0i}^2\right)$. Other choices of mean-zero residual are possible: see the discussion for log-linear SMMs in Clarke et al. (2015).

The form of the efficient instrument for the SVM is not straightforward but can be derived using semiparametric theory as sketched in Appendix SA.3. If ϵ_i, U_i and V_i are estimated simultaneously, the efficient instrument for each component residual is a linear combination of the expected derivatives of said residuals with respect to the joint-model parameters, where the scalar multipliers are functions of the moments of (ϵ_i, U_i, V_i). However, if we estimate U_i and V_i jointly but exclude the estimating equation for $\hat{\beta}$, the (locally) efficient instrument is

$$A_0(M_i) = E\left\{\begin{pmatrix} \frac{\partial U_i}{\partial \mu} & \frac{\partial V_i}{\partial \mu} \\ \frac{\partial U_i}{\partial \lambda} & \frac{\partial V_i}{\partial \lambda} \end{pmatrix}\middle| M_i\right\}\begin{pmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{pmatrix}^{-1},$$

where $\sigma_U^2 = E(U_i^2)$, $\sigma_V^2 = E(V_i^2)$ and $\sigma_{UV} = E(U_i V_i)$ (Newey 1993).

S3.4 The mode effect on the association between two continuous mixed-mode variables

For two mixed-mode continuous variables, the mode effect on the covariance can be estimated using the linear structural covariance model (SCM)

$$\text{cov}(X_i, Y_i | D_i, M_i) - \text{cov}(X_{i0}, Y_{0i} | D_i, M_i) = \sigma_1 D_i, \quad \text{(S3.10)}$$

subject to (S2.4) for X and for Y, and

$$E(X_{0i} Y_{0i} | M_i) = E(X_{0i} Y_{0i}). \quad \text{(S3.11)}$$

The target parameter is

$$\sigma_1 = \text{cov}(X_{1i}, Y_{1i} | D_i = 1) - \text{cov}(X_{i0}, Y_{0i} | D_i = 1)$$

with estimator

$$\begin{aligned} \hat{\sigma_1} = \frac{1}{\pi(1) - \pi(0)} [&E(X_i Y_i | M_i = 1) - E(X_i Y_i | M_i = 0) \\ &+ \pi(1)\{\mu_1^X \mu_1^Y \tau_1^X(1) \tau_1^Y(1) - \mu_1^X \tau_1^Y(1) - \mu_1^Y \tau_1^X(1)\} \\ &- \pi(0)\{\mu_1^X \mu_1^Y \tau_1^X(0) \tau_1^Y(0) - \mu_1^X \tau_1^Y(0) - \mu_1^Y \tau_1^X(0)\}] \end{aligned}$$

where μ_1^X and μ_1^Y are respectively the effects of mode on the means of X and Y under linear SMM (S2.3), $\tau_1^X(m) = E(X_i | D_i = 1, M_i = m)$ and $\tau_1^Y(m) = E(Y_i | D_i = 1, M_i = m)$ for $m = 0,1$.

For standard error estimation, the GMM residual satisfying $E(W_i | M_i) = E(W_i)$ is

$$W_i = \epsilon_i^X \epsilon_i^Y - D_i \sigma_1 + (U_i^X - \epsilon_i^X)(U_i^Y - \epsilon_i^Y), \quad \text{(S3.12)}$$

where $U_i^X = X_i - D_i \mu_1^X$ and $U_i^Y = Y_i - D_i \mu_1^Y$ are the residuals for SMM (S2.3) and $\epsilon_i^X = X_i - \beta_0^X - \beta_1^X M_i - \beta_2^X D_i - \beta_{12}^X M_i D_i$ and $\epsilon_i^Y = Y_i - \beta_0^Y - \beta_1^Y M_i - \beta_2^Y D_i - \beta_{12}^Y M_i D_i$ are the residuals of the association models, that is, the saturated linear regression of the survey variable on M_i, D_i and $M_i D_i$. Its zero-mean equivalent is thus

$$\overline{W}_i = \epsilon_i^X \epsilon_i^Y - \mu_{011} - D_i \sigma_1 + (U_i^X - \epsilon_i^X)(U_i^Y - \epsilon_i^Y), \quad \text{(S3.13)}$$

where $\mu_{011} = E(X_{0i} Y_{0i})$. The same comments regarding the form of the efficient instrument for the SVM in Appendix SA.3 are pertinent to the SCM.

S3.5 Adjusting for complex sampling designs

In general, if our (semi-)parametric model leads to $\mathbf{s}_i(\boldsymbol{\theta})$ such that $E\{\mathbf{s}_i(\boldsymbol{\theta})\} = \mathbf{0}$ and the dimensions of $\mathbf{s}_i(\boldsymbol{\theta})$ and $\boldsymbol{\theta}$ are equal, the method of moments finds $\widehat{\boldsymbol{\theta}} = \{\boldsymbol{\theta} : \mathbf{s}_n(\boldsymbol{\theta}) = \mathbf{0}\}$, where $\mathbf{s}_n(\widehat{\boldsymbol{\theta}}) = \sum_i \mathbf{s}_i(\widehat{\boldsymbol{\theta}})$ is the sum of these scores. The variance-covariance of $\widehat{\boldsymbol{\theta}}$ is estimated using the sandwich estimator

$$\widehat{V(\boldsymbol{\theta})} = \left\{\frac{\partial \mathbf{s}_n^T(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right\}^{-1} \left\{\sum_i \mathbf{s}_i(\widehat{\boldsymbol{\theta}}) \mathbf{s}_i^T(\widehat{\boldsymbol{\theta}})\right\} \left\{\frac{\partial \mathbf{s}_n(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^T}\right\}^{-1} = G^{-1} S G^{-T}, \quad \text{(S3.15)}$$

where $G^{-T} = (G^T)^{-1}$. The sandwich estimator is used rather than a plug-in estimator of the asymptotic variance, S^{-1}/n.

Stata has a suite of `svy` commands that can be used in conjunction with many of its standard estimation routines. However, these do not include the `gmm` command and so we implement the linearized version of (S3.15) used by `ivregress` among other routines. This simply replaces G by its weighted equivalent

$$G_w = \sum_i w_i \mathbf{s}_i(\widehat{\boldsymbol{\theta}}) \frac{\partial \mathbf{s}_i^T(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}},$$

where w_i is the survey weight for unit i. This is obtained by executing `gmm [pweight = h_indinub_xw]` etc. Then S is replaced by a design-consistent estimator of the variance of the sum $\sum_i \mathbf{s}_i(\widehat{\boldsymbol{\theta}})$, S_d. This can be implemented using `svy: total`. The linearized variance estimator is then

$$\widehat{V(\boldsymbol{\theta})}_{\text{lin}} = G_w^{-1} S_d G_w^{-T} \quad \text{(S3.16)}.$$

S4 Failure of the No Effect Modification (NEM) assumption

Consider the true linear SMM (S2.3) if NEM fails:

$$E(Y_i - Y_{0i}|D_i, M_i) = \mu_1(M_i)D_i, \quad \text{(S4.1)}$$

where $\mu_1(M_i) = E(Y_{1i} - Y_{0i}|D_i = 1, M_i)$.

This model is nonidentified if only the observed data are available because the resulting g-estimator would have three unknowns but only two equations. However, it could be identified if sample data were available on the same population from a survey where only face-to-face mode were used. In this case, with appropriate nonresponse adjustments, $\mu_0 = E(Y_{0i})$ could be estimated using these data rather than the observed data, leaving a g-estimator with two equations and two unknowns.

Two useful ways of expressing residual (S4.1) are

$$U_i = Y_{0i} - \mu_0 - (1 - M_i)D_i\mu_1(0) - M_iD_i\mu_1(1) = Y_{0i} - \mu_0 - D_i\mu_1(1) + (1 - M_i)D_i\Delta\mu_1, \quad \text{(S4.2)}$$

where $\Delta\mu_1 = \mu_1(1) - \mu_1(0)$. This trivially holds for the SMoM for $k \geq 1$.

Similarly, for log-linear SMM,

$$\begin{aligned} V_i &= Y_{0i}^2 \exp\{-(1 - M_i)D_i\lambda_2(0) - M_iD_i\lambda_2(1)\} - \nu_0 \\ &= Y_{0i}^2 \exp\{-D_i\lambda_2(1) - (1 - M_i)D_i\Delta\lambda_2\} - \nu_0, \quad \text{(S4.3)} \end{aligned}$$

where $\Delta\lambda_2 = \lambda_2(1) - \lambda_2(0)$.

The first parameterization of (S4.2) and (S4.3) is explicitly in terms of the two M-specific parameters from which the target parameters can be calculated as follows:

$$E(Y_{1i} - Y_{0i}|D_i = 1) = (1 - p)\mu_1(0) + p\mu_1(1) = \mu_1 + p\Delta\mu_1, \quad \text{(S4.4)}$$

$$\frac{\text{var}(Y_{1i}|D_i = 1)}{\text{var}(Y_{0i}|D_i = 1)} = \frac{S_1^2}{(1 - p)S_{01}^2 e^{-\lambda_1} + pS_{11}^2 e^{-\lambda_1 - \Delta\lambda_1} + A}, \quad \text{(S4.5)}$$

where $p = \Pr(M_i = 1|D_i = 1)$,

$$A = (1-p)p\{\Delta\mu_1^2 + (\bar{S}_{11} - \bar{S}_{01})^2 - 2(\bar{S}_{11} - \bar{S}_{01})\Delta\mu_1\},$$

$S_1^2 = \text{var}(Y_i|D_i = 1)$, $\bar{S}_{m1} = E(Y_{1i}|M_i = m, D_i = 1)$ and $S_{m1}^2 = \text{var}(Y_{1i}|M_i = m, D_i = 1)$.

Finally, for the additive SCM, if NEM fails then

$$\text{cov}(X_i, Y_i|M_i, D_i) - \text{cov}(X_{0i}, Y_{0i}|M_i, D_i) = D_i\sigma_1(M_i) = D_i\sigma_1 + M_iD_i\Delta\sigma_1, \quad \text{(S4.6)}$$

in which case

$$\text{cov}(X_i, Y_i|D_i = 1) - \text{cov}(X_{0i}, Y_{0i}|D_i = 1) = \sigma_1 + p\Delta\sigma_1 + B, \quad \text{(S4.7)}$$

where

$$B = (1-p)p\{\Delta\mu_1^X\Delta\mu_1^Y - (\bar{S}_{11}^Y - \bar{S}_{01}^Y)\Delta\mu_1^X - -(\bar{S}_{11}^X - \bar{S}_{01}^X)\Delta\mu_1^Y\},$$

and the X and Y superscripts indicate the obvious (based on the preceding discussion) parameters.

Standard errors for (S4.2), (S4.5) and (S4.7) can then be obtained post-estimation using the delta method.

S5 Simulation Study Design

A simulation study was carried out by generating data from the following simple model:

1. Generate the dichotomous instrumental variable $Z \sim \text{Bernoulli}(\pi_{\text{IV}})$.
2. Generate independent error terms $e_0^Y \sim N(0, \sigma_0^2)$ and $e_1^Y \sim N(0, \sigma_1^2)$.
3. Generate unobserved confounding variable $U \sim N(0,1)$.
4. Calculate the face-to-face mode potential outcomes $Y_0 = \beta_0 + \beta_2 U + e_0^Y$.
5. Calculate the web mode potential outcomes $Y_1 = \beta_0 + \beta_1 + \beta_2 U + e_0^Y + e_1^Y$.
6. Generate mode selection dependent on Z and U as follows:
   $D \sim \text{Bernoulli}[1/\{1 + \exp(-\alpha_0 - \alpha_1 Z - \alpha_2 U)\}]$.
7. Calculate the observed outcome $Y = (1-D)Y_0 + DY_1$.

The parameter values π_{IV}, σ_{X0}^2, σ_{X1}^2, α_0, α_1, α_2, β_0, β_1 and β_2 are set by the user. β_1 is the average of the simple mode effects $Y_1 - Y_0 = \beta_1 + e_1^Y$. β_2 determines the standard deviation of the unobserved confounding variable, and σ_{X1} is the standard deviation of the mode effect heterogeneity. The study looked at the sequence of g-estimators for increasing sample sizes $n = 10^2, 10^3, 10^4$ and 10^5 for the linear SMM $E(Y - Y_0|D, Z) = D\mu_1$, the log-linear SMoM $E(Y^2|D, Z)/E(Y_0^2|D, Z) = \exp(D\lambda_2)$, and SVM $\text{Var}(Y|D, Z)/\text{Var}(Y_0|D, Z) = \exp(D\upsilon_1)$. Monte Carlo estimation is used to calculate the true values of $\mu_1 = E(Y_1 - Y_0|D = 1)$, $\exp(\lambda_2) = E(Y_1^2|D = 1)/E(Y_0^2|D = 1)$ and $\exp(\upsilon_1) = \text{Var}(Y_1|D = 1)/\text{Var}(Y_0|D = 1)$ based on 10^7 draws.

Some results indicative of the performance of the g-estimators are displayed in Table S5.1 below.

Table S5.1 Example of g-estimator behaviour for three SMoMs

True values									
	π_{IV}	σ_0	σ_1	α_0	α_1	α_2	β_0	β_1	β_2
	0.5	2.5	1	0	1	2.5	5	2	1
Sample size		100		1000		10000		100000	
	F-statistic	**Avg.**	**SD**	**Avg.**	**SD**	**Avg.**	**SD**	**Avg.**	**SD**
		4.764	4.35	38.1	13	362	38.2	3615	127
Linear SMM ($\mu_1 = 2$)		**Est.**	**R. Bias**	**Est.**	**R. Bias**	**Est.**	**R. Bias**	**Est.**	**R. Bias**
MM- only	Coeff	1.294	-35.3%	1.955	-2.3%	2.005	0.3%	2.001	0.1%
	SE	95.9	381%	0.94	-3.0%	0.28	0.7%	0.09	2.3%
Augmented	Coeff	1.988	-0.6%	1.9999	0.0%	2.001	0.0%	2.002	0.1%
	SE	0.66	-0.6%	0.21	-1.0%	0.07	1.6%	0.02	0.0%
Log-linear SMoM ($\lambda_2 = 0.559$)									
MM-only	Coeff	0.627	12.4%	0.580	3.7%	0.575	2.8%	0.577	3.2%
	SE	1.83	-36.6%	0.29	-4.9%	0.08	0.0%	0.03	-3.7%
Aug (i)	Coeff	0.578	3.4%	0.564	0.8%	0.560	0.1%	0.562	0.5%
	SE	0.20	-0.5%	0.06	0.0%	0.02	0.0%	0.01	0.0%
Aug (ii)	Coeff	0.5805	3.8%	0.5644	0.9%	0.560	0.1%	0.561	0.3%
	SE	0.172	-4.4%	0.056	0.0%	0.02	1.2%	0.01	0.0%
SVM (υ_1 = 0.038)									
MM-only	Coeff	-0.112	-394%	-0.021	-157%	0.035	-7.1%	0.038	1.3%
	SE	10.8	645%	0.79	20.5%	0.16	1.3%	0.05	0.0%
Aug (i)	Coeff	0.068	80.2%	0.036	-4.0%	0.039	1.6%	0.039	2.1%
	SE	0.46	-2.1%	0.12	-0.9%	0.04	0.6%	0.01	0.0%
Aug (ii)	Coeff	0.092	144%	0.0447	17.9%	0.039	4.0%	0.038	1.1%
	SE	0.53	10.1%	0.107	-0.9%	0.03	0.0%	0.01	0.0%

Notes: R(elative) Bias = 100(Average – Truth)/abs(Truth); SD = Standard Deviation; SE = Standard Error; Aug(i)= mixed-mode data augmented by independent face-to-face sample fitted using just-identified one-step GMM; Aug (ii) = as Aug(i) except constants retained as instruments and fitted using two-step GMM.

The first set of results is for the g-estimator for μ_1 from the linear SMM for Y. In the first row, the estimator uses only the mixed-mode sample data and so is equivalent to 2SLS. The relative biases for the estimator of μ_1 and its standard error are both large for $n = 100$ but are considerably smaller, less than two percent, for $n = 1000$, and close to zero for $n = 10000$ and 100000. The large bias for $n = 100$ is explained by noting that Z is a weak instrument because the average F-statistic is $4.8 < 10$. In contrast, the average F-statistics all exceed 10 for the larger sample sizes, with an F-statistic of 362 for $n = 10000$, the sample size most relevant to the example in this paper.

The second row of Table S5.1 contains the results where the g-estimator is augmented by an unbiased estimator of $\mu_0 = E(Y_{0i})$ from an independent face-to-face-only sample also of size n. The relative bias is small (less than 2 percent) for all sample sizes and the standard errors smaller by a factor of roughly four.

The second set of results is for the g-estimator of λ_2 from the log-linear SMoM for Y^2. The same pattern of the results as for μ_1 is apparent; the main differences are that the relative biases for $n = 100$ are large but less large than for μ_1, but not close to zero for the larger sample sizes, despite being small. Additionally, there are two rows for the g-estimator augmented by an unbiased estimator of is augmented by an unbiased estimator of $\mu_{02} = E(Y_{0i}^2)$: the first is for the one-step GMM and the second for the two-step GMM. The difference between the two is that the one-step estimator dispenses with the constant instrument required to estimate λ_2 and μ_{02} freely, whereas the second one includes it in order to create an over-identified model with more instruments (1 and Z) than parameters. The two-step estimator is asymptotically efficient given the choice of instrumental variables and so potentially has smaller standard errors. However, the

results show that the choice made very little difference to the relative bias of the point or standard error estimates.

Finally, the third set of results is for the g-estimator of υ_1 from the log-linear SVM for Y. The results show that this g-estimator performs considerably less well than those above. The relative biases are large for $n = 100$ and 1000, even augmented by unbiased estimators of μ_0 and μ_{02}, but relative biases for $n = 10000$ and 100000 are nonzero but small.

S6 Limitations of the Indicator Method

In this section, we discuss the indicator method and its limitations when the outcome and predictor variables are subject to mode effects.

Take two variables Y_i and X_i and suppose that we wish to use ordinary least squares (OLS) to estimate the effect of mode on the regression coefficient of X_i in the regression of Y_i and X_i. The basic indicator method involves the simple linear regression of Y on X, D and the interaction term XD.

In the presence of nonrandom mode selection in an observational study, control variables $\mathbf{C}_i$ available can be used to adjust for selection effects provided that

$$D_i \perp\!\!\!\perp (Y_{0i}, X_{0i}, Y_{1i}, X_{0i}) | \mathbf{C}_i \quad \text{(S6.1)}$$

holds.

In this case, the effects of mode can be written $\tilde{\alpha}_1 - \tilde{\alpha}_0$ and $\tilde{\beta}_1 - \tilde{\beta}_0$, where

$$Y_{0i} = \tilde{\alpha}_0 + \tilde{\beta}_0 X_{0i} + \boldsymbol{\gamma}_0^T \mathbf{C}_i + e_{0i} \text{ and } Y_{1i} = \tilde{\alpha}_1 + \tilde{\beta}_1 X_{1i} + \boldsymbol{\gamma}_1^T \mathbf{C}_i + e_{1i}, \quad \text{(S6.2)}$$

and $E(e_{0i}|X_{0i}, \mathbf{C}_i) = E(e_{1i}|X_{1i}, \mathbf{C}_i) = 0$. The true mean does not have to be linear in X_i and $\mathbf{C}_i$ but if so then it must be recognised that the estimand of the adjusted OLS estimator is merely a measure of association rather than a parameter of the true model.

However, the indicator method is considerably more limited when analysing data from a sequential design where, as in this application, we do not think that (S6.1) is plausible. The IV indicator method is a 2SLS version of the basic indicator method above: stage one involves regressing D_i on M_i to obtain $E(D_i|M_i)$ and stage two regressing Y_i on X_i, $E(D_i|M_i)$ and $X_i E(D_i|M_i)$.

The mode effects are now $\alpha_1^{[1]} - \alpha_0^{[1]}$ and $\beta_1^{[1]} - \beta_0^{[1]}$, where

$$Y_{0i} = \alpha_0^{[1]} + \beta_0^{[1]} X_{0i} + e_{0i}^{[1]} \text{ and } Y_{1i} = \alpha_0^{[1]} + \beta_1^{[1]} X_{1i} + e_{1i}^{[1]}, \quad \text{(S6.3)}$$

subject to $E\left(e_{0i}^{[1]}|X_{0i}, D_i = 1\right) = E\left(e_{1i}^{[1]}|X_{1i}, D_i = 1\right) = 0$. These are the differences between the intercept and slopes using web and face-to-face data among those choose web. Note again that the true mean does not have to be linear.

We show below that one of the following assumptions is generally required to estimate $\beta_1^{[1]} - \beta_0^{[1]}$:

1. Non-random mode selection is independent of X_{0i}; or

2. The simple mode effect for X, $X_{0i} = X_{1i} + \varphi_i$, satisfies $E(\varphi_i|X_{1i} = x, D_i = 1) = 0$ and $E(Y_{0i}|X_{0i}, X_{1i} = x, D_i = 1) = E(Y_{0i}|X_{0i}, D_i = 1)$.

Both assumptions are strong. The first assumption is so strong that it is unnecessary to use IVs if it holds: simply regressing Y on X, D and interaction XD is unbiased and consistent for the mode effect.

The second assumption requires the mode effect to have mean zero and so unrelated to the observed value of X. While this might be plausible in some cases, it will not generally hold and is anyway impossible to verify.

However, a corollary of assumption 2 is that the IV indicator method can be used if X_i is a single-mode or mode-invariant variable because it trivially holds if either $X_{di} = X_i$ for all i (single-mode) or $X_{1i} = X_{0i}$ (mode-invariant) because both parts of assumption 2 would hold.

To demonstrate the role of assumption 1, expand the observed mean function as follows:

$$\begin{aligned} E(Y_i|X_i = x, M_i) &= \Pr(D_i = 0|M_i)\, E(Y_{0i}|X_{0i} = x, D_i = 0) + \Pr(D_i = 1|M_i)\, E(Y_{1i}|X_{1i} = xD_i = 1) \\ &= \alpha_0^{[0]} + \beta_0^{[0]}x + \left(\alpha_1^{[1]} - \alpha_0^{[0]}\right) E(D_i|M_i) + \left(\beta_1^{[1]} - \beta_0^{[0]}\right) xE(D_i|M_i), \end{aligned}$$

where $Y_{di} = \alpha_d^{[j]} + \beta_d^{[j]} X_{di} + e_{di}^{[j]}$ subject to $E\left(e_{di}^{[j]}|X_{di}, D_i = j\right) = 0$ are the simple linear regressions of Y_{di} on X_{di} among those choosing mode j $(d, j = 0{,}1)$.

The form of this model involves using the predicted choice of mode $E(D_i|M_i)$ rather than D_i itself, and includes the interaction between X_i and $E(D_i|M_i)$ as well as the main effects. However, this is no different to the naïve regression of Y_i on X_i and D_i, which similarly leads to a reduced-form regression model of the same form, that is,

$$\begin{aligned} E(Y_i|X_i = x, D_i) &= (1 - D_i)E(Y_{0i}|X_{0i} = x, D_i = 0) + D_i E(Y_{1i}|X_{1i} = xD_i = 1) \\ &= \alpha_0^{[0]} + \beta_0^{[0]}x + \left(\alpha_1^{[1]} - \alpha_0^{[0]}\right) D_i + \left(\beta_1^{[1]} - \beta_0^{[0]}\right) xD_i. \end{aligned}$$

In neither case is the coefficient of the interaction term generally equal to $\beta_1^{[1]} - \beta_0^{[1]}$, that is, the effect of mode on the regression coefficient among those choosing web. Only if $\beta_0^{[0]} = \beta_0^{[1]} = \beta_0$, which holds only under the implausible assumption that selection is independent of X_{0i}, does the indicator method return a meaningful mode effect.

To show the role of assumption 2, expand

$$\begin{aligned} E(Y_i|X_i = x, M_i) &= E\{Y_{0i} + D_i(Y_{1i} - Y_{0i})|X_i = x, M_i\} \\ &= E(Y_{0i}|X_i = x) + \Pr(D_i = 1|M_i)\, E(Y_{1i} - Y_{0i}|X_i = x, D_i = 1) \\ &= \alpha_0 + \beta_0 x + E(Y_{1i} - Y_{0i}|X_i = x, D_i = 1)E(D_i|M_i). \end{aligned}$$

This approach requires the following assumptions about the mode measurement model for X_i:

$$\begin{aligned} E(Y_{1i} - Y_{0i}|X_i = x, D_i = 1) &= E(Y_{1i} - Y_{0i}|X_{1i} = x, D_i = 1) \\ &= \alpha_1^{[1]} + \beta_1^{[1]}x - E(Y_{0i}|X_{1i} = x, D_i = 1), \end{aligned}$$

where

$$E(Y_{0i}|X_{1i} = x, D_i = 1) = E_{X_{i0}|X_{1i}=x,D_i=1}\{E(Y_{0i}|X_{0i}, X_{1i} = x, D_i = 1)\}.$$

Progress can be made if the inner expectation on the right-hand side satisfies

$$E(Y_{0i}|X_{0i}, X_{1i} = x, D_i = 1) = E(Y_{0i}|X_{0i}, D_i = 1),$$

so that

$$E_{X_{i0}|X_{1i}=x,D_i=1}\{E(Y_{0i}|X_{0i},D_i=1)\}=\alpha_0^{[1]}+\beta_0^{[1]}E(X_{0i}|X_{1i}=x,D_i=1).$$

Furthermore, if the mode measurement model connecting the two outcomes, $X_{0i}=X_{1i}+\varphi_i$, satisfies $E(X_{1i}+\varphi_i|X_{1i}=x,D_i=1)=x$, that is, the average mode effect $E(\varphi_i|X_{1i}=x,D_i=1)=0$ for all values of X_{1i}, then

$$E(Y_{0i}|X_{1i}=x,D_i=1)=\alpha_0^{[1]}+\beta_0^{[1]}x,$$

and

$$E(Y_{1i}-Y_{0i}|X_i=x,D_i=1)=\alpha_1^{[1]}+\beta_1^{[1]}x-\alpha_0^{[1]}-\beta_0^{[1]}x,$$

as required.

From this, it follows that the required mode effect is obtained if $X_i=X_{0i}$ or $X_i=X_{1i}$ for all i (single-mode variable) or $X_{1i}=X_{0i}$ (mode-invariant variable).

S7 Mode Effect on the Maximum Likelihood Estimator

S7.1 Methodological development

Let $\hat{\theta}_0=\hat{\theta}(Y_0)$ be the maximum likelihood estimator (MLE) which would have been obtained had everyone used face-to-face mode, and $\hat{\theta}=\hat{\theta}(Y)$ be the MLE obtained using the observed mixed-modes data.

Now let θ^* be the probability limit of $\hat{\theta}$, where

$$\bar{s}\big(\hat{\theta};Y\big)=n^{-1}\sum_{i=1}^{n}s\big(\hat{\theta};Y_i\big)=(1-\hat{\pi})\bar{s}_0\big(\hat{\theta};Y\big)+\hat{\pi}\bar{s}_1\big(\hat{\theta};Y\big)=0,\quad\text{(S7.1)}$$

is the sample analogue of the population moment restriction $E\{s(\theta^*;Y)\}=0$, $\bar{s}_d\big(\hat{\theta};Y\big)=\sum_{i=1}^{n}s\big(\hat{\theta};Y_i\big)I(D_i=d)/\sum_{i=1}^{n}I(D_i=d)$ is the mean (observed-data) score among those who choose mode $d=0,1$, and $\hat{\pi}=n^{-1}\sum_{i=1}^{n}I(D_i=1)$ is unbiased for $\pi=\Pr(D_i=1)$.

The usual first-order Taylor series expansion of $\bar{s}\big(\hat{\theta};Y\big)$ around θ^* gives $\bar{s}\big(\hat{\theta};Y\big)=\bar{s}(\theta^*;Y)-\mathcal{F}_0(\theta^*)\big(\hat{\theta}-\theta^*\big)=0$ under (S7.1) so that

$$\bar{s}(\theta^*;Y)=\mathcal{F}(\theta^*)\big(\hat{\theta}-\theta^*\big),\quad\text{(S7.2)}$$

where $\mathcal{F}(\theta^*)=\mathrm{E}\{-\partial s^T(\theta;Y_i)/\partial\theta|_{\theta=\theta^*}\}$ is the Fisher information (ignoring $o_p(1)$ terms).

Likewise, $\hat{\theta}_0$ is the solution to

$$\bar{s}\big(\hat{\theta}_0;Y_0\big)=n^{-1}\sum_{i=1}^{n}s\big(\hat{\theta}_0;Y_{0i}\big)=(1-\hat{\pi})\bar{s}_0\big(\hat{\theta}_0;Y_0\big)+\hat{\pi}\bar{s}_1\big(\hat{\theta}_0;Y_0\big)=0,\quad\text{(S7.3)}$$

where $\bar{s}_d\big(\hat{\theta}_0;Y_0\big)=\sum_{i=1}^{n}s\big(\hat{\theta}_0;Y_{0i}\big)I(D_i=d)/\sum_{i=1}^{n}I(D_i=d)$ is the mean (face-to-face) score among those who choose mode $d=0,1$.

For a sequence (in n) of $\hat{\theta}_0$ lying in a neighbourhood of θ^*, a first-order Taylor series expansion of (S7.3) around θ^* gives $\bar{s}\big(\hat{\theta}_0;Y_0\big)=\bar{s}(\theta^*;Y_0)-\mathcal{F}_0(\theta^*)\big(\hat{\theta}_0-\theta^*\big)=0$ so that

$$\bar{s}(\theta^*;Y_0)=\mathcal{F}_0(\theta^*)\big(\hat{\theta}_0-\theta^*\big),\quad\text{(S7.4)}$$

where $\mathcal{F}_0(\theta^*) = \mathrm{E}\{-\partial s^T(\theta; Y_{0i})/\partial\theta|_{\theta=\theta^*}\}$. Note that $\mathcal{F}_0$ is not now the expected Fisher information matrix because θ^* is not the probability limit of $\hat{\theta}_0$.

Under the convenient approximation

$$\mathcal{F}_0(\theta^*) \approx \mathcal{F}(\theta^*), \quad \text{(S7.5)}$$

it follows from combining (S7.1) and (S7.3) that

$$\hat{\theta} - \hat{\theta}_0 \approx V(\theta^*)\{\bar{s}(\theta^*; Y) - \bar{s}(\theta^*; Y_0)\} = \hat{\pi}V(\theta^*)\{\bar{s}_1(\theta^*; Y) - \bar{s}_1(\theta^*; Y_0)\}, \quad \text{(S7.6)}$$

where $V(\theta^*) = \mathcal{F}^{-1}(\theta^*)$.

To evaluate the asymptotic distribution of $\hat{\theta} - \hat{\theta}_0$, first define sample average $\Delta\bar{s}_1 = \bar{s}_1(\theta^*; Y) - \bar{s}_1(\theta^*; Y_0)$, $\Delta s_i = s(\theta^*; Y_i) - s(\theta^*; Y_{0i})$ and $\Delta s_1 = \mathrm{E}(\Delta s_i | D_i = 1)$. Then rewrite (S7.6) as

$$\hat{\theta} - \left(\hat{\theta}_0 + \pi V(\theta^*)\Delta s_1\right) \approx V(\theta^*)(\hat{\pi}\,\Delta\bar{s}_1 - p\Delta s_1), \quad \text{(S7.7)}$$

and apply the (multivariate) central limit theorem to the random variable $D_i\Delta s_i$ (because $\hat{\pi}\,\Delta\bar{s}_1 = n^{-1}\sum_{i=1}^{n} D_i\Delta s_i$ with mean $p\Delta s_1 \neq 0$). If we write

$$\hat{\pi}\,\Delta\bar{s}_1 - \pi\Delta s_1 = \{n^{-1}Q_1(\theta^*)\}^{\frac{1}{2}}\{n^{-1}Q_1(\theta^*)\}^{-\frac{1}{2}}(\hat{\pi}\,\Delta\bar{s}_1 - \pi\Delta s_1)$$

then (S7.7) can be rewritten as

$$\sqrt{n}\left\{\hat{\theta} - \left(\hat{\theta}_0 + \pi V(\theta^*)\Delta s_1\right)\right\} \approx V(\theta^*)Q_1^{1/2}(\theta^*)\left[\{n^{-1}Q_1(\theta^*)\}^{-1/2}(\hat{\pi}\,\Delta\bar{s}_1 - \pi\Delta s_1)\right],$$

where

$$Q_1(\theta^*) = \mathrm{var}(D_i\Delta s_i) = E\{(D_i\Delta s_i - \pi\Delta s_1)(D_i\Delta s_i - \pi\Delta s_1)^T\} = \pi\mathrm{E}\left(\Delta s_i\Delta s_i^T \middle| D_i = 1\right) - \pi^2\Delta s_1\Delta s_1^T$$

is the variance-covariance of zero-mean $D_i\Delta s_i - \pi\Delta s_1$.

It follows from applying the (multivariate) central limit theorem that $\{n^{-1}Q_1(\theta^*)\}^{-1/2}(\hat{\pi}\,\Delta\bar{s}_1 - \pi\Delta s_1) \sim \mathcal{N}(0, I)$ so that

$$\hat{\theta} - \hat{\theta}_0 \dot{\sim} \mathcal{N}\{\pi V(\theta^*)\Delta s_1, n^{-1}V(\theta^*)Q_1(\theta^*)V(\theta^*)\} \quad \text{(S7.8)}$$

as $n \to \infty$.

Equation (S7.6) simplifies as follows when the density of Y_i is a member of the curved exponential family of distributions such that

$$f(y_i; \theta) = h(y_i)\exp\{\eta^T(\theta)T(y_i) - A(\theta)\},$$

where $T(.)$ is the sufficient statistic for natural parameter η. In this case,

$$\bar{s}(\theta; Y) = \frac{\partial\eta^T(\theta)}{\partial\theta}\left\{(1 - \hat{\pi})\bar{T}_0(Y) + \hat{\pi}\bar{T}_1(Y) - \frac{\partial A(\eta)}{\partial\eta}\right\},$$

where $\bar{T}_d(Y) = \sum_{i=1}^{n} I(D_i = d)T(Y_i)/\sum_{i=1}^{n} I(D_i = d)$, which gives

$$\hat{\theta} - \hat{\theta}_0 = \hat{\pi}V(\theta^*)\frac{\partial\eta^T(\theta^*)}{\partial\theta}\{\bar{T}_1(Y) - \bar{T}_1(Y_0)\} \quad \text{(S7.9)}$$

and $\bar{T}_1(Y) - \bar{T}_1(Y_0)$ can be estimated by fitting the multivariate SMM

$$E\{T(Y_i) - T(Y_{0i})|D_i, M_i\} = D_i\mu_{1T},$$

so that

$$\hat{\theta} - \hat{\theta}_0 = \hat{\pi} V\left(\hat{\theta}\right) \frac{\partial \eta^T\left(\hat{\theta}\right)}{\partial \theta} \hat{\mu}_{1T}. \quad \text{(S7.10)}$$

Alternatively, we can directly estimate the face-to-face score function as follows: rewrite (S7.3) as

$$\bar{s}(\theta; Y_0) = n^{-1} \sum\nolimits_i \; (1 - D_i) \frac{\partial \eta^T}{\partial \theta} T(Y_i) - \frac{\partial A}{\partial \theta} + \frac{\partial \eta^T}{\partial \theta} \left\{ n^{-1} \sum\nolimits_i D_i \, T(Y_{0i}) \right\}.$$

The final parenthesised term is consistent for $\pi E\{T(Y_{0i}) | D_i = 1\} = \pi E\{T(Y_i) - \mu_T | D_i = 1\}$, where μ_T is the parameter of the linear SMoM $E\{T(Y_i) - T(Y_{0i}) | D_i, M_i\} = D_i \mu_T$. Plugging this into the expression above and simplifying yields the plug-in estimator

$$\widehat{\bar{s}(\theta; Y_0)} = n^{-1} \sum\nolimits_i \; \frac{\partial \eta^T}{\partial \theta} \{T(Y_i) - D_i \hat{\mu}_T\} - \frac{\partial A}{\partial \theta}, \quad \text{(S7.11)}$$

where $\hat{\theta}_0 = \left\{\theta : \widehat{\bar{s}(\theta; Y_0)} = 0\right\}$.

In terms of variance estimation, we cannot rely on the limiting distribution (S7.8) because Q_1 cannot be estimated, and we have derived no such convenient expression for (S7.11). Submitted Paper (2020) suggest using the bootstrap to estimate the standard errors but base significance on a (joint) hypothesis test of $\mu_{1T} = 0$ based on weighed linearized variance estimation of the SMoM.

Example S7.1: Consider the normal linear regression model for outcome Y with two predictors X_1 and X_2 where all three variables are mean-centred. The mode of interest is

$$Y_i = b_1 X_{1i} + b_2 X_{2i} + e_i, \quad \text{(S7.12)}$$

where residual e_i is assumed to be normally distributed with variance σ_e^2. If none of the three variables is mode-invariant, we must jointly model all three. We take the joint distribution to be normal with zero mean and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_X & \sigma_{XY} \\ \sigma_{XY}^T & \sigma_Y^2 \end{pmatrix},$$

where $\sigma_{XY}^T = (\sigma_{1Y}, \sigma_{2Y})$ are respectively the covariances between Y and X_1 and Y and X_2,

$$\Sigma_X = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ & \sigma_2^2 \end{pmatrix}$$

is the variance-covariance matrix of X_1 and X_2, and σ_Y^2 is the variance of the outcome.

The connection between the natural parameters of the exponential-family representation of this distribution and $\theta = (\sigma_1^2, \sigma_{12}, \sigma_2^2, b_1, b_2, \sigma_e^2)^T$ is given by standard results:

$$\begin{pmatrix} \eta_1 & \eta_2 & \eta_3 \\ & \eta_4 & \eta_5 \\ & & \eta_6 \end{pmatrix} = \Sigma^{-1} = \begin{pmatrix} \Sigma_X^{-1} + b b^T / \sigma_e^2 & -b/\sigma_e^2 \\ -b^T/\sigma_e^2 & 1/\sigma_e^2 \end{pmatrix},$$

where $b^T = (b_1, b_2)$ and $\sigma_e^2 = \sigma_Y^2 - \sigma_{XY}^T b$ are respectively the coefficients and residual variance of (S7.11). The sufficient statistics for $\eta_1 = \sigma_2^2/d + b_1^2/\sigma_e^2$, $\eta_2 = -\sigma_{12}/d + b_1 b_2/\sigma_e^2$, $\eta_4 = \sigma_1^2/d + b_2^2/\sigma_e^2$, $\eta_3 = -b_1/\sigma_e^2$, $\eta_5 = -b_2/\sigma_e^2$ and $\eta_6 = 1/\sigma_e^2$ are

respectively $T_1 = X_1^2, T_2 = X_1X_2, T_4 = X_2^2, T_3 = X_1Y, T_5 = X_2Y$ and $T_6 = Y^2$ (where $d = \det(\Sigma_X)$).

None of the derivatives in $\partial\eta^T/\partial\theta$ taken with respect to σ_1^2, σ_{12} or σ_2^2 depend on b_1, b_2 or σ_e^2 so we need only focus on the derivatives taken with respect to the parameters of (S7.12). These are given as follows:

$$\frac{\partial\eta_1}{\partial b_1} = \frac{2b_1}{\sigma_e^2}, \frac{\partial\eta_1}{\partial b_2} = 0, \frac{\partial\eta_1}{\partial\sigma_e^2} = -\frac{b_1^2}{\sigma_e^4},$$

$$\frac{\partial\eta_2}{\partial b_1} = \frac{b_2}{\sigma_e^2}, \frac{\partial\eta_2}{\partial b_2} = \frac{b_1}{\sigma_e^2}, \frac{\partial\eta_2}{\partial\sigma_e^2} = -\frac{b_1b_2}{\sigma_e^4},$$

$$\frac{\partial\eta_4}{\partial b_1} = 0, \frac{\partial\eta_4}{\partial b_2} = \frac{2b_2}{\sigma_e^2}, \frac{\partial\eta_4}{\partial\sigma_e^2} = -\frac{b_2^2}{\sigma_e^4},$$

$$\frac{\partial\eta_3}{\partial b_1} = -\frac{1}{\sigma_e^2}, \frac{\partial\eta_3}{\partial b_2} = 0, \frac{\partial\eta_3}{\partial\sigma_e^2} = \frac{b_1}{\sigma_e^4},$$

$$\frac{\partial\eta_5}{\partial b_1} = 0, \frac{\partial\eta_5}{\partial b_2} = -\frac{1}{\sigma_e^2}, \frac{\partial\eta_5}{\partial\sigma_e^2} = \frac{b_2}{\sigma_e^4},$$

and

$$\frac{\partial\eta_6}{\partial b_1} = \frac{\partial\eta_6}{\partial b_2} = 0, \frac{\partial\eta_6}{\partial\sigma_e^2} = -\frac{1}{\sigma_e^4}.$$

In this situation, equation (S7.6) for $\theta = (b_1, b_2, \sigma_e^2)^T$ is

$$\hat{\theta} - \hat{\theta}_0 \approx \hat{\pi}V(\hat{\theta})\{\bar{s}_1(\hat{\theta}; Y) - \bar{s}_1(\hat{\theta}; Y_0)\} = \hat{\pi}V(\hat{\theta})\sum_{k=1}^{6}\left.\frac{\partial\eta_k}{\partial\theta}\right|_{\theta=\hat{\theta}}\hat{\mu}_{T1k},$$

where

$$\mu_{T1k} = E\{T_k(Z_i) - T_k(Z_{0i})|D_i = 1\}$$

is estimated using the appropriate SMoM, $Z_i = (X_{1i}, X_{2i}, Y_i)^T$ with its face-to-face equivalent Z_{0i}, and is the observed-data estimated variance covariance matrix.

Example S7.2: Now consider estimating the effect of mode on generalized linear models of the form

$$g\{E(Y_i|X_i)\} = X_i^T\beta,$$

where g is the link function and Y_i (given X_i) is a member of the over-dispersed exponential family with a density function of the form

$$f(y_i) = h(y_i, \sigma)\exp\left(\frac{\eta_i T(y_i) - A(\eta_i)}{\sigma}\right),$$

where σ is the over-dispersion scale parameter.

We focus on simple models like the normal, exponential, Bernoulli and Poisson where $T(y_i) = y_i$ and σ is a known constant. For these cases, the efficient score is

$$\bar{s}(\beta) = \frac{1}{n}\sum_i \frac{\partial\eta_i}{\partial\beta}\left\{\frac{Y_i - g^{-1}(X_i^T\beta)}{\sigma}\right\}.$$

In contrast to the case above, the natural parameter η_i varies between individuals through the implicit conditioning on X_i so neither (S7.6) nor (S7.11) are applicable because both assume independent and identically distributed realizations when only the first of these now holds.

For example, the logistic regression model assumes that Y_i is Bernoulli distributed and logit link $g(p) = \log\{p/(1-p)\}$ so that $\eta_i = x_i^T\beta$, $T(y_i) = y_i$, $A(\eta_i) = \log\{1+\exp(x_i^T\beta)\}$, $\sigma = 1$, and

$$\bar{s}(\beta) = n^{-1}\sum_i X_i\left[Y_i - 1/\{1+\exp(-X_i^T\beta)\}\right]$$

is the score function based on the mixed-modes data.

We can see that the form of this canonical model is loglinear in the sufficient statistic for β, Y_iX_i, but that $1/\{1+\exp(-X_i^T\beta)\}$ is not. Thus, the same approach as before can be used to adjust Y_iX_i based on the SMoM

$$E(Y_iX_i - Y_{0i}X_{0i}|D_i, M_i) = D_i\psi_{yx}.$$

The same approach can also be used for the non-loglinear term but only by proceeding iteratively as follows. Let $\beta^{(k)}$ be the current estimate of the regression parameter, which can be treated as a constant in the update step. Now estimate the mode effect on $X_iZ_i^{(k)}$ using the SMoM

$$E\left(X_iZ_i^{(k)} - X_{0i}Z_{0i}^{(k)}\middle| D_i, M_i\right) = D_i\tau_{yx}^{(k)},$$

where $Z_i^{(k)} = 1/\{1+\exp(-X_i^T\beta^{(k)})\}$ and $Z_{0i}^{(k)} = 1/\{1+\exp(-X_{0i}^T\beta^{(k)})\}$. The estimated score function at iteration k is thus

$$\bar{s}^{(k)}(\beta) = n^{-1}\sum_i X_iY_i - X_iZ_i - D_i\left(\hat{\psi}_{xy} - \tau_{yx}^{(k)}\right),$$

where $Z_i = \log\{1+\exp(X_i^T\beta)\}$, and $\beta^{(k+1)} = \{\beta: \bar{s}^{(k)}(\beta) = 0\}$.

A similar approach can be used straightforwardly for other canonical generalized linear models. For the non-canonical models like the probit for binary outcomes with link $g(p) = \Phi(p)$, SMoMs are required for $Y_i\,\partial\Phi(X_i^T\beta)/\partial\beta$ and $\Phi^{-1}(X_i^T\beta)\,\partial\Phi(X_i^T\beta)/\partial\beta$ and both must be iteratively estimated as β is updated.

S7.2 Simulation Study

To demonstrate the performance of the estimators based on the approaches outlined above, we carried out a simulation study to estimate the mode effect on the association between Y and X when both variables are subject to mode effects.

The first example concerns the effect of mode of the coefficient of X in the linear regression of Y on X when both variables are continuous. The data were generated as follows:

1. Generate the dichotomous instrumental variable $Z \sim \text{Bernoulli}(\pi_{\text{IV}})$.
2. Generate independent error terms $e_0^X \sim N(0, \sigma_0^2)$ and $e_0^Y \sim N(0, \sigma_0^2)$.

3. Generate correlated heterogeneity terms $\begin{pmatrix} e_1^X \\ e_1^Y \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1^2 \\ & \sigma_1^2 \end{pmatrix}\right\}$.
4. Generate unobserved confounding variable $U \sim N(0,1)$.
5. Calculate the face-to-face mode potential outcomes $X_0 = \beta_0 + \beta_2 U + e_0^X$ and $Y_0 = \beta_0 + \beta_2 U + e_0^Y$.
6. Calculate the web mode potential outcomes $X_1 = \beta_0 + \beta_1 + \beta_2 U + e_0^X + e_1^X$ and $Y_1 = \beta_0 + \beta_1 + \beta_2 U + e_0^Y + e_1^Y$.
7. Generate mode selection dependent on Z and U as follows: $D \sim \text{Bernoulli}[1/\{1 + \exp(-\alpha_0 - \alpha_1 Z - \alpha_2 U)\}]$.
8. Calculate the observed outcome $Y = (1 - D)Y_0 + DY_1$.

The approximate method (S7.10) is a direct estimate of the mode effect; the exact method is the difference between the observed association and the solution to the mode-effect-adjusted score equation (S7.11).

The bias and spread of the approximate and exact methods for an indicative example are shown in the table below. The population parameters result, respectively, in mixed-mode and face-to-face regression models

$$Y = 0.0004 + 0.4608X + e \text{ and } Y_0 = 0.0005 + 0.5005X_0 + e,$$

with a small mode effect on the regression coefficient: –0.0397.

The second example is the logistic regression of binary Y on X. The data-generating procedure above is modified as follows to yield a binary outcome: $Y_d \equiv I(Y_d > 0)$. then the resulting logistic regression models are

$$\text{logit}\Pr(Y = 1|X) = -0.0251 + 0.5857X \text{ and } \text{logit}\Pr(Y_0 = 1|X_0) = 0.0022 + 0.6750X_0$$

corresponding to a mode effect of 0.0893. The focus of the study is the use of exact method for generalised linear models so the target parameter is 0.6750 the 'slope' coefficient of X_0.

The results using the exact approach for both linear and logistic regression are in line with those for the SMoMs in that the relative bias is less than 10% for sample sizes of 10,000 and 100,000. This is expected because the method is driven by the bias of the SMoM estimates. The estimator is also accurate in that its standard deviation is small enough to permit a statistical test to produce a significant finding from a hypothesis test.

Table S7.1 Estimated mode effects on coefficients of linear and logistic regression using SMoM method

True values									
π_{IV}	σ_0	σ_1	ρ	α_0	α_1	α_2	β_0	β_1	β_2
0.5	1	0.5	–0.1	0	1	2.5	0	0	1
Sample size		100		1000		10000		100000	
Linear regression		**Est.**	**R. Bias**	**Est.**	**R. Bias**	**Est.**	**R. Bias**	**Est.**	**R. Bias**
					True mode effect = –0.0397				
Approx	Mode effect	0.2238	+663%	0.0137	+134%	–0.0339	14.7%	–0.0372	6.4%
	SD	1.062		0.275		0.081		0.025	
	Converged	99%		100%		100%		100%	
Exact	Mode effect	0.0144	+63.8%	–0.0372	+6.5%	–0.0378	+4.8%	–0.0398	-0.3%
	SD	0.558		0.346		0.090		0.027	
	Converged	46%		90%		100%		100%	
Logistic Regression									
					True slope = 0.6750				
Exact	Slope	0.5408	-19.9%	0.8189	+21.3%	0.6928	+2.7%	0.6718	-0.5%
	SD	0.869		1.110		0.208		0.055	
	Converged	19%		84%		100%		100%	

Notes: R(elative) Bias = 100(Average – Truth)/Abs(Truth); SD = Standard Deviation; SE = Standard Error;

The results of this study also indicate that an advantage of the approximate method for small sample sizes is that it almost always converges, but this is offset by large bias for sample size of 100. However, for the larger sample size, it outperforms the exact method in terms of mean square error $(0.0137 + 0.0397)^2 + 0.275^2 = 0.078$ to $(-0.0372 + 0.0397)^2 + 0.346^2 = 0.120$ as well as in terms of convergence (100% to 90%). A complete assessment of the relative strengths of these two approaches would be an interesting topic for further investigation.

S8 Adjusting for non-response

S8.1 Inverse probability weighting

If R_i^{cc} is the complete-cases response indicator and we have control variables $\mathbf{C}_i$ satisfying

$$R_i^{cc} \perp\!\!\!\perp \begin{pmatrix} Y_{0i} \\ Y_{1i} \\ M_i \\ D_i \end{pmatrix} \Bigg| \mathbf{C}_i, \quad \text{(S8.1)}$$

then the inverse probability weighted (IPW) estimator for SMM

$$E(Y_i - Y_{0i}|D_i, M_i) = D_i\mu_1, \quad \text{(S8.2)}$$

is simply (S2.9) weighted by $w(\mathbf{C}_i, M_i, D_i) = 1/\Pr(R_i^{cc} = 1|\mathbf{C}_i, M_i, D_i)$, that is,

$$\sum\nolimits_i w(\mathbf{C}_i, M_i, D_i)\mathbf{a}_0(M_i)\overline{U}_i R_i^{cc} = \mathbf{0}. \quad \text{(S8.3)}$$

However, if the control variables are related to the mode effect on the mean so that

$$E(Y_i - Y_{0i}|D_i, M_i, \mathbf{C}_i) = D_i\mu_1(\mathbf{C}_i), \quad \text{(S8.4)}$$

SMM (S8.2) cannot satisfy NEM because

$$E(Y_i - Y_{0i}|D_i, M_i) = E\{E(Y_i - Y_{0i}|D_i, M_i, \mathbf{C}_i)|D_i, M_i\} = D_i E\{\mu_1(\mathbf{C}_i)|D_i = 1, M_i\}$$
$$\neq D_i E\{\mu_1(\mathbf{C}_i)|D_i = 1\},$$

under either of the (control variable-conditional) data generating processes defined in Supplementary Information S.1, unless $D_i \perp\!\!\!\perp M_i|Y_i^*, \mathbf{C}_i$ or $D_i \perp\!\!\!\perp M_i|Y_i^*, Y_{0i}, Y_{1i}, \mathbf{C}_i$, that is, the control variables also control for the dependence of D on M in the data generating process for mode selection, where Y_i^* is the latent true value of the survey characteristics.

S8.2 Linear SMMs

Under (S8.1), we have that

$$E(Y_i - Y_{0i}|D_i, M_i, \mathbf{C}_i, R_i^{cc} = 1) = E(Y_i - Y_{0i}|D_i, M_i, \mathbf{C}_i).$$

where the right-hand side follows SMM (S8.4) and $\mu_1(\mathbf{C}_i) = E(Y_{1i} - Y_{0i}|D_i = 1, \mathbf{C}_i)$.

The most obvious way to handle this dependence would be to specify a parametric SMM for $\mu_1(\mathbf{C}_i)$ but, for simplicity, we would prefer to estimate $\mu_1 = E\{\mu_1(\mathbf{C}_i)|D_i = 1\}$ as if there were no association between the mode effects and control variables, that is, using

$$E(Y_i - Y_{0i}|D_i, M_i, \mathbf{C}_i) = D_i\mu_1, \quad \text{(S8.5)}$$

despite NEM not being satisfied. To show that the naïve g-estimator based on (S8.5) is not consistent for μ_1, consider the residual one would use if $\mathbf{C}_i$ and mode effect were unrelated, that is,

$$U_i = Y_i - D_i\mu_1^*,$$

where $\mu_1^* \neq \mu_1$. A locally efficient but naive g-estimator for μ_1^* is the solution to

$$\sigma_U^{-2} \sum_i \{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}U_i = 0,$$

which, ignoring σ_U^{-2}, is consistent for the population moment restriction

$$E[\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}U_i] = 0. \quad \text{(S8.6)}$$

Now rewrite

$$U_i = Y_i - D_i\mu_1(\mathbf{C}_i) + D_i\{\mu_1(\mathbf{C}_i) - \mu_1^*\} = \overline{U}_i + D_i\{\mu_1(\mathbf{C}_i) - \mu_1^*\},$$

where $\overline{U}_i$ satisfies CMI $E(\overline{U}_i|M_i, \mathbf{C}_i) = E(\overline{U}_i|\mathbf{C}_i)$. Hence, equation (S8.6) can be rewritten

$$\begin{aligned}
&E[\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\overline{U}_i] \\
&\quad + E[\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\{\mu_1(\mathbf{C}_i) - \mu_1^*\}D_i] \\
&\quad = E[\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}E(\overline{U}_i|M_i, \mathbf{C}_i)] \\
&\quad + E[\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\{\mu_1(\mathbf{C}_i) - \mu_1^*\}D_i] \\
&\quad = E[E\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)|\mathbf{C}_i\}E(\overline{U}_i|\mathbf{C}_i)] \\
&\quad + E[\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\{\mu_1(\mathbf{C}_i) - \mu_1^*\}D_i] \\
&\quad = 0 + E[\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\{\mu_1(\mathbf{C}_i) - \mu_1^*\}D_i] = 0,
\end{aligned}$$

from which equality to zero gives

$$\mu_1^* = \frac{E[\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\mu_1(\mathbf{C}_i)|D_i = 1]}{E\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)|D_i = 1\}} \neq \mu_1.$$

However, if (S8.6) had been weighted by

$$w_i(\mathbf{C}_i) = \frac{1}{E\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\,|D_i = 1, \mathbf{C}_i\}},$$

then

$$\begin{aligned}
&E[w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}U_i] \\
&\quad = E[w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}E(\overline{U}_i|\mathbf{C}_i)] \\
&\quad + \Pr(D_i = 1)\,E[w_i(\mathbf{C}_i)E\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)|\mathbf{C}_i\}\{\mu_1(\mathbf{C}_i) - \mu_1\}|D_i \\
&\quad = 1] \\
&\quad = E[w_i(\mathbf{C}_i)E\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)|\mathbf{C}_i\}E(\overline{U}_i|\mathbf{C}_i)] \\
&\quad + \Pr(D_i = 1)\,E[\{\mu_1(\mathbf{C}_i) - \mu_1\}|D_i = 1]\Pr(D_i = 1) \\
&\quad = \Pr(D_i = 1)\,[E\{\mu_1(\mathbf{C}_i)|D_i = 1\} - \mu_1] = 0, \quad \text{(S8.7)}
\end{aligned}$$

where one should note μ_1^* has been replaced by μ_1 so it follows that the weighted g-estimator is consistent for μ_1.

To improve the estimated standard errors, we modify the above argument to work with the mean-zero residual

$$U_i = Y_i - E(Y_i|\mathbf{C}_i) - \{D_i - \Pr(D_i = 1|\mathbf{C}_i)\}\mu_1 = \overline{Y}_i - \overline{D}_i\mu_1,$$

where $\overline{Y}_i = Y_i - E(Y_i|\mathbf{C}_i)$ and $\overline{D}_i = D_i - \Pr(D_i = 1|\mathbf{C}_i)$ require the analyst to specify and fit additional models for $E(Y_i|\mathbf{C}_i)$ and $\Pr(D_i = 1|\mathbf{C}_i)$. However, the same arguments as above based around

$$U_i = \overline{Y}_i - \overline{D}_i\mu_1(\mathbf{C}_i) + \overline{D}_i\{\mu_1(\mathbf{C}_i) - \mu_1^*\} = \overline{U}_i + \overline{D}_i\{\mu_1(\mathbf{C}_i) - \mu_1^*\}$$

give

$$\sum\nolimits_i w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}U_i = 0, \quad \text{(S8.8)}$$

because CMI $E(\overline{U}_i|M_i, \mathbf{C}_i) = 0$. This is also a consistent estimating equation for μ_1 because

$$\begin{aligned}
&E[w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}U_i] \\
&\quad = E[w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}E(\overline{U}_i|M_i, \mathbf{C}_i)] \\
&\quad + E[w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\{\mu_1(\mathbf{C}_i) - \mu_1\}D_i^*] \\
&\quad = 0 + E[w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\{\mu_1(\mathbf{C}_i) - \mu_1\}D_i] \\
&\quad + E[w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\{\mu_1(\mathbf{C}_i) - \mu_1\}\Pr(D_i = 1|\mathbf{C}_i)] \\
&\quad = E[w_i(\mathbf{C}_i)\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)\}\{\mu_1(\mathbf{C}_i) - \mu_1\}D_i] \\
&\quad + E[w_i(\mathbf{C}_i)E\{\Pr(D_i = 1|M_i, \mathbf{C}_i) - \Pr(D_i = 1|\mathbf{C}_i)|\mathbf{C}_i\}\{\mu_1(\mathbf{C}_i) - \mu_1\}\Pr(D_i = 1|\mathbf{C}_i)] \\
&\quad = 0 + E[\{\mu_1(\mathbf{C}_i) - \mu_1\}D_i] = 0.
\end{aligned}$$

S4.3 Log-linear SMMs

The same weighting strategy does not work for log-linear SMoMs. To show why, consider the same set up as above for the log-linear SMoM

$$\begin{aligned}
&\log\{E(Y_i^2|D_i, M_i, \mathbf{C}_i, R_i^{cc} = 1)\} - \log\{E(Y_{0i}^2|D_i, M_i, \mathbf{C}_i, R_i^{cc} = 1)\} \\
&\qquad = \log\{E(Y_i^2|D_i, M_i, \mathbf{C}_i)\} - \log\{E(Y_{0i}^2|D_i, M_i, \mathbf{C}_i)\} = D_i\mu_2(\mathbf{C}_i), \quad \text{(S8.9)}
\end{aligned}$$

where $\exp\{\mu_2(\mathbf{C}_i)\} = E(Y_{1i}^2|D_i = 1, \mathbf{C}_i)/E(Y_{0i}^2|D_i = 1, \mathbf{C}_i)$. Together with CMI

$$E(Y_{0i}^2|M_i, \mathbf{C}_i) = E(Y_{0i}^2|M_i, \mathbf{C}_i), \quad \text{(S8.10)}$$

this model implies that

$$V_i = Y_i^2 \exp\{-\mu_2(\mathbf{C}_i)D_i\}, \quad \text{(S8.11)}$$

satisfies CMI assumption

$$E(V_i|M_i, \mathbf{C}_i) = E(V_i|\mathbf{C}_i), \quad \text{(S8.12)}$$

whereas the simpler residual

$$V_i^* = Y_i^2 \exp(-\mu_2^* D_i), \quad \text{(S8.13)}$$

derived from

$$\log\{E(Y_i^2|D_i, M_i, \mathbf{C}_i)\} - \log\{E(Y_{0i}^2|D_i, M_i, \mathbf{C}_i)\} = \mu_2^* D_i \quad \text{(S8.14)}$$

does not satisfy CMI $E(V_i^*|M_i, \mathbf{C}_i) = E(V_i^*|\mathbf{C}_i)$.

Finally, note that the target parameter is

$$\exp(\mu_2) = \frac{E(Y_{1i}^2|D_i = 1)}{E(Y_{0i}^2|D_i = 1)} = \frac{E(Y_i^2|D_i = 1)}{E[Y_i^2 \exp\{-\mu_2(\mathbf{C}_i)\}|D_i = 1]}. \quad \text{(S8.15)}$$

but the g-estimator is consistent for

$$\exp(\mu_2^*) = \frac{E\{b_0(M_i, \mathbf{C}_i)Y_i^2 D_i\}}{E[b_0(M_i, \mathbf{C}_i)Y_i^2 D_i \exp\{-\mu_2(\mathbf{C}_i)\}]} \neq \exp(\mu_2),$$

where $b_0(M_i, \mathbf{C}_i)$ is the efficient instrument.

However, the dependence of (S8.15) on Y_i^2 ultimately prevents any choice of $w_i(\mathbf{C}_i)$ correcting for this bias. For example, consider expanding (S8.13) as

$$V_i^* = V_i + [\exp(-\mu_2^*) - \exp\{-\mu_2(\mathbf{C}_i)\}]Y_i^2 D_i,$$

in which case the general form of the g-estimator for (S8.14) is

$$E\{b_0(M_i, \mathbf{C}_i)V_i^*\} = E\{b_0(M_i, \mathbf{C}_i)Y_i^2 D_i\}\exp(-\mu_2^*) - E[b_0(M_i, \mathbf{C}_i)Y_i^2 D_i \exp\{-\mu_2(\mathbf{C}_i)\}] = 0,$$

where

$$b_0(M_i, \mathbf{C}_i) \propto E(D_i Y_i^2|M_i, \mathbf{C}_i) - E(D_i Y_i^2|\mathbf{C}_i),$$

is a locally efficient choice satisfying $E\{b_0(M_i, \mathbf{C}_i)|\mathbf{C}_i\} = 0$. Then

$$\begin{aligned} &E\{w_i(\mathbf{C}_i)b_0(M_i, \mathbf{C}_i)V_i^*\} \\ &\quad = \Pr(D_i = 1)\, E\{w_i(\mathbf{C}_i)b_0(M_i, \mathbf{C}_i)Y_i^2|D_i = 1\}\exp(-\mu_2^*) \\ &\quad - \Pr(D_i = 1)\, E[w_i(\mathbf{C}_i)b_0(M_i, \mathbf{C}_i)Y_i^2 \exp\{-\mu_2(\mathbf{C}_i)\}|D_i = 1] = 0, \end{aligned}$$

but almost surely

$$E\{w_i(\mathbf{C}_i)b_0(M_i, \mathbf{C}_i)E(Y_i^2|D_i = 1, M_i, \mathbf{C}_i)|D_i = 1\} \neq E(Y_i^2|D_i = 1),$$

$$E[w_i(\mathbf{C}_i)b_0(M_i, \mathbf{C}_i)E(Y_i^2|D_i = 1, M_i, \mathbf{C}_i)\exp\{-\mu_2(\mathbf{C}_i)\}|D_i = 1] \neq E[Y_i^2 \exp\{-\mu_2(\mathbf{C}_i)\}|D_i = 1]$$

for any choice of weight. The best that could be done would be to choose

$$w_i(\mathbf{C}_i) = 1/E\{b_0(M_i, \mathbf{C}_i)Y_i^2|D_i = 1, \mathbf{C}_i\},$$

in which case the g-estimator would be consistent for

$$\exp(\mu_2^*) = \frac{1}{E[\exp\{-\mu_2(\mathbf{C}_i)\}|D_i = 1]} = \left[E\left\{\frac{E\left(Y_{0i}^2|D_i = 1, \mathbf{C}_i\right)}{E\left(Y_{1i}^2|D_i = 1, \mathbf{C}_i\right)}\middle| D_i = 1\right\}\right]^{-1},$$

but while this does not depend on b_0 it would not target parameter (S8.15).

■ Appendices:

SA1 Analytical expression for SVM estimator

To show this, use the law of iterated expectations to obtain

$$E\left(Y_{0i}^2|M_i\right) = E\left\{E\left(Y_{0i}^2|D, M_i\right)|M_i\right\}.$$

Now convert the inner expectation into a variance plus remainder term as follows:

$$E\left(Y_{0i}^2|D, M_i\right) = \text{var}(Y_{0i}|D, M_i) + E^2(Y_{0i}|D, M_i).$$

This expression can now be written in terms of linear and log-linear SMM parameters as follows:

$$\begin{aligned} E\left(Y_{0i}^2|D, M_i\right) &= \exp(-D\lambda)\text{var}(Y_i|D, M_i) + \{E(Y_i|D, M_i) - D\mu\}^2 \\ &= \{(1 - D) + D\exp(-\lambda)\}\text{var}(Y_i|D, M_i) + \{E(Y_i|D, M_i) - D\mu\}^2. \end{aligned}$$

We further use that

$$\begin{aligned} E\left(Y_{0i}^2|M_i\right) &= \Pr(D = 0|M_i)\, E\left(Y_{0i}^2|D = 0, M_i\right) + \Pr(D = 1|M_i)\, E\left(Y_{0i}^2|D = 1, M_i\right) \\ &= \Pr(D = 0|M_i)\,[\text{var}(Y_i|D = 0, M_i) + \{E(Y_i|D = 0, M_i)\}^2] \\ &+ \Pr(D = 1|M_i)\,[\exp(-\lambda)\text{var}(Y_i|D = 1, M_i) + \{E(Y_i|D = 1, M_i) - \mu\}^2]. \end{aligned}$$

Then solve $E\left(Y_{0i}^2|M_i = 0\right) = E\left(Y_{0i}^2|M_i = 1\right)$ by expanding out the left and right-hand sides as

$$\begin{aligned} &\Pr(D = 0|M_i = 0)\, E\left(Y_{0i}^2|D = 0, M_i = 0\right) + \Pr(D = 1|M_i = 0)\, E\left(Y_{0i}^2|D = 1, M_i = 0\right) \\ &= \Pr(D = 0|M_i = 0)\,[\text{var}(Y_i|D = M_i = 0) + \{E(Y_i|D = M_i = 0)\}^2] \\ &+ \Pr(D = 1|M_i = 0)\,[\exp(-\lambda)\text{var}(Y_i|D = 1, M_i = 0) \\ &+ \{E(Y_i|D = 1, M_i = 0) - \mu\}^2] \\ &= \Pr(D = 0|M_i = 1)\, E\left(Y_{0i}^2|D = 0, M_i = 1\right) \\ &+ \Pr(D = 1|M_i = 1)\, E\left(Y_{0i}^2|D = M_i = 1\right) \\ &= \Pr(D = 0|M_i = 1)\,[\text{var}(Y_i|D = 0, M_i = 1) + \{E(Y_i|D = 0, M_i = 1)\}^2] \\ &+ \Pr(D = 1|M_i = 1)\,[\exp(-\lambda)\text{var}(Y_i|D = M_i = 1) \\ &+ \{E(Y_i|D = M_i = 1) - \mu\}^2] \end{aligned}$$

to obtain a closed-form expression for $\exp(\lambda)$.

SA2 Analytical expression for SCM estimator

To show this, use the law of iterated expectations to obtain

$$E(X_{0i}Y_{0i}|M_i) = E\{E(X_{0i}Y_{0i}|D, M_i)|M_i\},$$

and then convert the inner expectation into a covariance plus remainder term as follows:

$$E(X_{0i}Y_{0i}|D,M_i) = \mathrm{cov}(X_{0i},Y_{0i}|D,M_i) + E\{E(X_{0i}|D,M_i)E(Y_{0i}|D,M_i)|M_i\}.$$

This expression can now be written like so

$$E(X_{0i}Y_{0i}|D,M_i) = \mathrm{cov}(X_i,Y_i|D,M_i) - D\sigma^{XY} + \{E(X_i|D,M_i) - D\mu^X\}\{E(Y_i|D,M_i) - D\mu^Y\}.$$

Now we use that

$$\begin{aligned} E(X_{0i}Y_{0i}|M_i) &= \Pr(D=0|M_i)\,E(X_{0i}Y_{0i}|D=0,M_i) + \Pr(D=1|M_i)\,E(X_{0i}Y_{0i}|D=1,M_i) \\ &= \Pr(D=0|M_i)\,[\mathrm{cov}(X_i,Y_i|D=0,M_i) + E(X_i|D=0,M_i)E(Y_i|D=0,M_i)] \\ &+ \Pr(D=1|M_i)\,[\mathrm{cov}(X_i,Y_i|D=1,M_i) - \sigma^{XY} \\ &+ \{E(X_i|D=1,M_i) - \mu^X\}\{E(Y_i|D=1,M_i) - \mu^Y\}]. \end{aligned}$$

Aside: Before proceeding, we recall that the law of iterated expectations for covariances can be written as

$$\mathrm{cov}(X,Y|V) = E_W\{\mathrm{cov}(X,Y|W,V)|V\} + \mathrm{cov}_W\{E(X|W,V),E(Y|W,V)|V\}.$$

In other words, the conditional covariance (given V) equals the sum of the average conditional covariance (given (V,W)) and the covariance of the conditional averages (both given (V,W)). The second of these terms can also be written as

$$\begin{aligned} \mathrm{cov}_W\{E(X|W,V),E(Y|W,V)|V\} &= E_W[\{E(X|W,V) - E(X|V)\}\{E(Y|W,V) - E(Y|V)\}] \\ &= E_W\{E(X|W,V)E(Y|W,V)\} - E(X|V)E(Y|V). \quad \text{(S3.15)} \end{aligned}$$

Proceeding line by line:

$$\begin{aligned} &\Pr(D=0|M_i)\,\mathrm{cov}(X_i,Y_i|D=0,M_i) + \Pr(D=1|M_i)\,\mathrm{cov}(X_i,Y_i|D=1,M_i) \\ &\quad = E_D\{\mathrm{cov}(X_i,Y_i|D,M_i)|M_i\}. \end{aligned}$$

$$\begin{aligned} &\Pr(D=0|M_i)\,E(X_i|D=0,M_i)E(Y_i|D=0,M_i) \\ &\quad + \Pr(D=1|M_i)\,E(X_i|D=0,M_i)E(Y_i|D=0,M_i) \\ &\quad = E_D\{E(X_i|D,M_i)E(Y_i|D,M_i)|M\}. \end{aligned}$$

$$-\Pr(D=1|M_i)\,\sigma^{XY}.$$

$$\Pr(D=1|M_i)\,\{\mu^X\mu^Y - E(X_i|D=1,M_i)\mu^Y - E(Y_i|D=1,M_i)\mu^X\}.$$

The first two of these can be written in terms of $\mathrm{cov}(X_i,Y_i|M_i)$ using (S3.15) as follows:

$$E_D\{\mathrm{cov}(X_i,Y_i|D,M_i)|M_i\} + E_D\{E(X_i|D,M_i)E(Y_i|D,M_i)|M\} = \mathrm{cov}(X_i,Y_i|M_i) + E(X_i|M_i)E(Y_i|M_i),$$

and the square completed to give

$$\begin{aligned} &\Pr(D=1|M_i)\,\{\mu^X\mu^Y - E(X_i|D=1,M_i)\mu^Y - E(Y_i|D=1,M_i)\mu^X\} \\ &\quad = \Pr(D=1|M_i)\,\{\mu^X - E(X_i|D=1,M_i)\}\{\mu^Y - E(Y_i|D=1,M_i)\} \\ &\quad - \Pr(D=1|M_i)\,E(X_i|D=1,M_i)E(Y_i|D=1,M_i). \end{aligned}$$

Hence,

$$\begin{aligned} E(X_{0i}Y_{0i}|M_i) &= \mathrm{cov}(X_i,Y_i|M_i) + E(X_i|M_i)E(Y_i|M_i) - \Pr(D=1|M_i)\,\sigma^{XY} \\ &+ \Pr(D=1|M_i)\,\{\mu^X - E(X_i|D=1,M_i)\}\{\mu^Y - E(Y_i|D=1,M_i)\} \\ &- \Pr(D=1|M_i)\,E(X_i|D=1,M_i)E(Y_i|D=1,M_i). \end{aligned}$$

Finally, expanding $E(X_{0i}Y_{0i}|M_i=0)=E(X_{0i}Y_{0i}|M_i=1)$ out as

$$\begin{aligned}\operatorname{cov}(X_i,Y_i|M_i=0)+E(X_i|M_i=0)E(Y_i|M_i=0)-\Pr(D=1|M_i=0)\,\sigma^{XY}\\ +\Pr(D=1|M_i=0)\,\{\mu^X-E(X_i|D=1,M_i=0)\}\{\mu^Y-E(Y_i|D=1,M_i=0)\}\\ -\Pr(D=1|M_i=0)\,E(X_i|D=1,M_i=0)E(Y_i|D=1,M_i=0)\\ =\operatorname{cov}(X_i,Y_i|M_i=1)+E(X_i|M_i=1)E(Y_i|M_i=1)-\Pr(D=1|M_i=1)\,\sigma^{XY}\\ +\Pr(D=1|M_i=1)\,\{\mu^X-E(X_i|D=M_i=1)\}\{\mu^Y-E(Y_i|D=M_i=1)\}\\ -\Pr(D=1|M_i=1)\,E(X_i|D=M_i=1)E(Y_i|D=M_i=1).\end{aligned}$$

gives the closed-form expression for σ^{XY}.

SA3 Sketch of derivation of efficient instrument for SVM estimator

Recall that the model residuals are

$$\epsilon=Y-\beta_0-\beta_1M-\beta_2D-\beta_{12}MD,\;\; U=Y-\mu_0-\mu_1D,$$

and

$$\begin{aligned}V&=e^{-\lambda_1M}(Y-\beta_0-\beta_1M-\beta_2D-\beta_{12}MD)^2+\{\beta_0+\beta_1M+(\beta_2-\mu_1)D+\beta_{12}MD\}^2-\lambda_0\\&=e^{-\lambda_1M}\epsilon^2+(U+\mu_0-\epsilon)^2-\lambda_0.\end{aligned}$$

all of which satisfy $E(\epsilon|M,D)=E(U|M)=E(V|M)=0$. The sketch follows Bowden and Vansteelandt (2011) and Tsiatis (2006, Theorem 1).

Note $\sigma_\epsilon^2(M,D)=E(\epsilon^2|M,D)$, $\sigma_{\epsilon U}(M)=E(\epsilon U|M)$, $\sigma_{\epsilon V}(M)=E(\epsilon V|M)$, $\sigma_U^2=E(U^2)$, $\sigma_V^2=E(V^2)$ $\sigma_{UV}=E(UV)$, and so on with dependence on M and D suppressed or explicitly excluded.

The form of the estimating equation is

$$\mathbf{g}\begin{pmatrix}\beta\\ \mu\\ \lambda\end{pmatrix}=\begin{Bmatrix}a_\beta(M,D)\\ a_\mu(M,D)\\ a_\lambda(M,D)\end{Bmatrix}\epsilon+\begin{Bmatrix}b_\beta(M)\\ b_\mu(M)\\ b_\lambda(M)\end{Bmatrix}U+\begin{Bmatrix}c_\beta(M)\\ c_\mu(M)\\ c_\lambda(M)\end{Bmatrix}V.$$

The parameterization β, μ and λ of the first two moments of $Y,Y_0|D=1,M$ is variation-independent so the semiparametrically efficient choices of a, b and c must satisfy

$$\begin{aligned}&E\left[\left\{\begin{pmatrix}a_\beta\\ a_\mu\\ a_\lambda\end{pmatrix}\epsilon+\begin{pmatrix}b_\beta\\ b_\mu\\ b_\lambda\end{pmatrix}U+\begin{pmatrix}c_\beta\\ c_\mu\\ c_\lambda\end{pmatrix}V\right\}^T\begin{pmatrix}h_\beta\\ h_\mu\\ h_\lambda\end{pmatrix}\right]\\&\qquad=E\left[\left\{\begin{pmatrix}a_\beta\\ a_\mu\\ a_\lambda\end{pmatrix}\epsilon+\begin{pmatrix}b_\beta\\ b_\mu\\ b_\lambda\end{pmatrix}U+\begin{pmatrix}c_\beta\\ c_\mu\\ c_\lambda\end{pmatrix}V\right\}^T\left\{\begin{pmatrix}a_\beta\\ a_\mu\\ a_\lambda\end{pmatrix}\epsilon+\begin{pmatrix}b_\beta\\ b_\mu\\ b_\lambda\end{pmatrix}U+\begin{pmatrix}c_\beta\\ c_\mu\\ c_\lambda\end{pmatrix}V\right\}\right],\end{aligned}$$

for all h in the 8-dimensional Hilbert space of mean zero random variables satisfying $\|E(hh^T)\|<\infty$.

To determine the efficient choices, we begin by matching the left-hand and right-hand sides as follows:

$$a_\beta^T\colon E\big(h_\beta\epsilon\big|M,D\big)=\sigma_\epsilon^2a_\beta(M,D)+\sigma_{\epsilon U}b_\beta(M)+\sigma_{\epsilon V}c_\beta(M),\qquad\text{(SA3.1)}$$

$$b_\beta^T\colon E\big(h_\beta U\big|M\big)=E\big\{\sigma_{\epsilon U}a_\beta(M,D)\big|M\big\}+\sigma_U^2b_\beta(M)+\sigma_{UV}c_\beta(M),$$

$$c_\beta^T: E\big(h_\beta V\big|M\big) = E\big\{\sigma_{\epsilon V} a_\beta(M,D)\big|M\big\} + \sigma_{UV} b_\beta(M) + \sigma_V^2 c_\beta(M),$$

where (SA3.1) also implies that

$$E\big(h_\beta \epsilon\big|M\big) = E\big\{\sigma_\epsilon^2 a_\beta(M,D)\big|M\big\} + \sigma_{\epsilon U} b_\beta(M) + \sigma_{\epsilon V} c_\beta(M). \quad \text{(SA3.2)}$$

It also follows from constraints $E(\epsilon|M,D) = E(U|M) = E(V|M) = 0$ that

$$\frac{\partial}{\partial\beta} E(\epsilon|M,D) = E\big(h_\beta \epsilon\big|M,D\big) + E\left(\frac{\partial \epsilon}{\partial \beta}\middle| M,D\right) = 0 \Rightarrow E\big(h_\beta \epsilon\big|M,D\big) = E\left(-\frac{\partial \epsilon}{\partial \beta}\middle| M,D\right),$$

$$\frac{\partial}{\partial\beta} E(U|M) = E\big(h_\beta U\big|M\big) + E\left(\frac{\partial U}{\partial \beta}\middle| M\right) = 0 \Rightarrow E\big(h_\beta U\big|M\big) = 0,$$

$$\frac{\partial}{\partial\beta} E(V|M) = E\big(h_\beta V\big|M\big) + E\left(\frac{\partial V}{\partial \beta}\middle| M\right) = 0 \Rightarrow E\big(h_\beta V\big|M\big) = E\left(-\frac{\partial V}{\partial \beta}\middle| M\right).$$

Hence,

$$E\left(-\frac{\partial \epsilon}{\partial \beta}\middle| M,D\right) = \sigma_\epsilon^2 a_\beta(M,D) + \sigma_{\epsilon U} b_\beta(M) + \sigma_{\epsilon V} c_\beta(M), \quad \text{(SA3.3)}$$

$$0 = E\big\{\sigma_{\epsilon U} a_\beta(M,D)\big|M\big\} + \sigma_U^2 b_\beta(M) + \sigma_{UV} c_\beta(M), \quad \text{(SA3.4)}$$

$$E\left(-\frac{\partial V}{\partial \beta}\middle| M\right) = E\big\{\sigma_{\epsilon V} a_\beta(M,D)\big|M\big\} + \sigma_{UV} b_\beta(M) + \sigma_V^2 c_\beta(M). \quad \text{(SA3.5)}$$

To construct a closed-form solution, it is necessary to make the local assumption

$$\sigma_\epsilon^2(M,D) = \sigma_\epsilon^2(M), \sigma_{\epsilon U}(M,D) = \sigma_{\epsilon U}(M), \sigma_{\epsilon V}(M,D) = \sigma_{\epsilon V}(M),$$

under which (SA3.2-SA3.5) become

$$E\left(-\frac{\partial \epsilon}{\partial \beta}\middle| M,D\right) = \sigma_\epsilon^2 a_\beta(M,D) + \sigma_{\epsilon U} b_\beta(M) + \sigma_{\epsilon V} c_\beta(M), \quad \text{(SA3.6)}$$

$$0 = E\big\{\sigma_{\epsilon U} a_\beta(M,D)\big|M\big\} + \sigma_U^2 b_\beta(M) + \sigma_{UV} c_\beta(M), \quad \text{(SA3.7)}$$

$$E\left(-\frac{\partial V}{\partial \beta}\middle| M\right) = E\big\{\sigma_{\epsilon V} a_\beta(M,D)\big|M\big\} + \sigma_{UV} b_\beta(M) + \sigma_V^2 c_\beta(M), \quad \text{(SA3.8)}$$

and (SA3.6) further implies

$$E\left(-\frac{\partial \epsilon}{\partial \beta}\middle| M\right) = \sigma_\epsilon^2 E\big\{a_\beta(M,D)\big|M\big\} + \sigma_{\epsilon U} b_\beta(M) + \sigma_{\epsilon V} c_\beta(M). \quad \text{(SA3.9)}$$

Then from (SA3.7),

$$E\big\{a_\beta(M,D)\big|M\big\} = -\frac{1}{\sigma_{\epsilon U}}\big\{\sigma_U^2 b_\beta(M) + \sigma_{UV} c_\beta(M)\big\},$$

which can be substituted into (SA3.9) to give

$$b_\beta(M) = \frac{1}{A}\left\{E\left(-\frac{\partial \epsilon}{\partial \beta}\middle| M\right) - B c_\beta(M)\right\}. \quad \text{(SA3.10)}$$

Now (SA3.8) can similarly be written

$$E\left(-\frac{\partial V}{\partial \beta}\middle| M\right) = Cb_\beta(M) + Dc_\beta(M),$$

and substituted into (SA3.10) to give

$$c_\beta(M) = \frac{A}{AD-CB}\left\{E\left(-\frac{\partial V}{\partial \beta}\middle| M\right) - \frac{C}{A}E\left(-\frac{\partial \epsilon}{\partial \beta}\middle| M\right)\right\}, \quad \text{(SA3.11)}$$

and hence

$$b_\beta(M) = \frac{1}{A}\left\{\left(1-\frac{C}{A}\right)E\left(-\frac{\partial \epsilon}{\partial \beta}\middle| M\right) - \frac{AB}{AD-CB}E\left(-\frac{\partial V}{\partial \beta}\middle| M\right)\right\}. \quad \text{(SA3.12)}$$

These can then be substituted via (SA3.6) to give

$$a_\beta(M,D) = \frac{1}{\sigma_\epsilon^2}\left\{E\left(-\frac{\partial \epsilon}{\partial \beta}\middle| R,M\right) - \sigma_{\epsilon U}b_\beta(M) - \sigma_{\epsilon V}c_\beta(M)\right\}. \quad \text{(SA3.13)}$$

Note that the constants A, B, C and D are

$$A = \sigma_{\epsilon U} - \frac{\sigma_\epsilon^2\sigma_U^2}{\sigma_{\epsilon U}} = \sigma_{\epsilon U}\left(1-\rho_{\epsilon,U}^{-2}\right), B = \sigma_{\epsilon V} - \frac{\sigma_\epsilon^2\sigma_{UV}}{\sigma_{\epsilon U}}, C = \sigma_{UV} - \frac{\sigma_U^2\sigma_{\epsilon V}}{\sigma_{\epsilon U}} \text{ and } D = \sigma_V^2 - \frac{\sigma_\epsilon^2\sigma_{\epsilon V}}{\sigma_{\epsilon U}},$$

where $\rho_{\epsilon,U}$ is the Pearson correlation coefficient between ϵ and U.

A similar derivation is gives closed-form expressions for the remaining terms.

References

Bowden, J. and Vansteelandt, S. (2011) Mendelian randomization analysis of case-control data using structural mean models. *Stat. Med.* **30**(6), 678-694.

Clarke, P.S., Palmer, T.M. and Windmeijer, F. (2015) Estimating structural mean models with multiple instrumental variables using the generalized method of moments. *Stat. Sci.* **30**(1), 96-117.

Submitted Paper (2020) Estimating mode effects from a sequential mixed-mode experiment using structural moment models.

Hernán, M.A. and Robins, J.M. (2006) Instruments for causal inference: an epidemiologist's dream? *Epidemiol.* **17**(4), 360-372.

Liang, K.-Y. and Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **74**(1), 13-22.

Newey, W.K. (1993) Efficient estimation of models with conditional moment restrictions. Chapt. 16 in *Handbook of Statistics* Vol. 11 (eds. G.S. Maddala, C.R. Rao and H.D. Vinod). North-Holland: Amsterdam, pp. 419-454.

Tsiatis, A.T. (2006) *Semiparametric Theory and Missing Data*. London: Springer.