Review: RRBB-2024-0076 Quantifying Potential Selection Bias in Observational Research: Simulations and Analyses Exploring Religion and Depression Using a Prospective UK Cohort Study (ALSPAC)

Joseph A. Bulbulia

2024-04-02

Decision

Major revisions required before considering for publication.

Summary

This submission addresses the critical issue of sample and target population mismatch, or "selection bias." The authors specifically focus on the implications of panel attrition in a longitudinal study of pregnant women from Bristol in the early 1990s, examining the bias introduced when estimating the effects of religiosity on depression. The paper employs simulation and comparison of different datasets to evaluate problems of mismatch.

Despite its merits, the manuscript contains methodological flaws that I believe preclude publication in its current form. However, with substantial revisions addressing the concerns outlined below, the study has the potential for publication.

Concerns and Recommendations

1. **Methodological problems**: the study suffers from significant methodological issues. These issues must be thoroughly addressed in a revision.
2. **Misuse of regression coefficients**: The authors rely on regression coefficients to evaluate threats to external validity, but this is misguided. This approach fails to account for treatment effect modification, leading to inaccurate assessments of bias. Below, I demonstrate the problem using a simulation.
3. **Focus on marginal effect estimates**: a revised version of the paper should include a more robust analysis that accounts for the complexities of sample/target population mismatch and its implications estimating causal effects They should include interaction terms when simulating data, and they should specifically evaluate whether differences in the distribution of effect-modifiers in the sample and target population lead to differences between sample and target population causal effect estimates. Anything less is arguably unrealistic for longitudinal research in the study of religion.
4. **Encouragement**: despite the current shortcomings, the submission addresses an issue of considerable importance and widespread confusion. The authors should be encouraged to undertake the necessary revisions and resubmit their work. Appropriately considered findings will make a valuable contribution to the field.
5. **Acknowledgment of fallibility**: I offer my review with the hope that it will assist the authors in strengthening their paper. I apologise if I have overlooked important details in their work.

Simulation to clarify the issues

To understand why we must focus on marginal effect estimates, consider the following simulation and analysis.

First, we load the `stdReg` library, which obtains marginal effect estimates by simulating counterfactuals under different levels of treatment (Sjölander 2016). If a treatment is continuous, the levels can be specified.

We also load the `parameters` library, which creates nice tables (Lüdecke *et al.* 2020).

```
# to obtain marginal effects
library(stdReg)
# to create nice tables
library(parameters)
```

Next, we write a function to simulate data for the sample and and target populations.

We assume the treatment effect is the same in the sample and target population. We will assume that the coefficient for the effect-modifier and the coefficient for interaction are the same. We assume no unmeasured confounding throughout the study. We assume only selective attrition of one effect modifier such that the baseline population differs from the sample population at the end of the study.

That is: **the distribution of effect modifiers is the only respect in which the sample will differ from the target population.**

This function will generate data under a range of scenarios.[1]

```
# function to generate data for the sample and population,
# along with precise sample weights for the population, there are differences
# in the distribution of the true effect modifier but no differences in the treatment effect
# or the effect modification.all that differs between the sample and the population is
# the distribution of effect-modifiers.


# reproducability
set.seed(123)


# function based on margot::simulate_ate_data_with_weights,
# see: https://go-bayes.github.io/margot/

simulate_ate_data_with_weights <- function(n_sample = 10000, # sample n
                                           #population n
                                           n_population = 100000,
                                           # prob of effect modifier in sample
                                           p_z_sample = 0.1,
                                           #prob of effect modifier in the population
                                           p_z_population = 0.5,
                                           # treatment effect
                                           beta_a = 1,
                                           # coef of intervention
                                           beta_z = 2.5,
                                           # coef  of effect-modifier
                                           beta_az = 0.5,
```

[1] See documentation in the `margot` package: Bulbulia (2024)

```
                                        # effect modification of a by z
                                        noise_sd = .5) {

  # create sample data
  z_sample <- rbinom(n_sample, 1, p_z_sample) # simulate data for sample z
  a_sample <- rbinom(n_sample, 1, 0.5) # for sample treatment

  # simulate outcome
  y_sample <- beta_a * a_sample + beta_z * z_sample + beta_az * (a_sample * z_sample) +
    rnorm(n_sample, mean = 0, sd = noise_sd)  # use noise_sd for the noise term

  # put sample data in data frame
  sample_data <- data.frame(y_sample, a_sample, z_sample)

  # simulate population data, where the distribution of effect modifiers differs, but the treatment effect is th
  z_population <- rbinom(n_population, 1, p_z_population)
  a_population <- rbinom(n_population, 1, 0.5) # same effect of a on y
  y_population <- beta_a * a_population + beta_z * z_population +
    beta_az * (a_population * z_population) + rnorm(n_population, mean = 0, sd = noise_sd)  # noise

  # put population data in dataframe
  population_data <- data.frame(y_population, a_population, z_population)

  # simulate weighting based on z distribution difference
  weight_z_1 = p_z_population / p_z_sample # adjust weight for Z=1
  weight_z_0 = (1 - p_z_population) / (1 - p_z_sample) # adjust weight for Z=0
  weights <- ifelse(z_sample == 1, weight_z_1, weight_z_0)

  # add weights to sample_data
  sample_data$weights = weights

  # return list of data frames and weights
  list(sample_data = sample_data, population_data = population_data)
}

# simulate the data -- you can use different parameters
data <- simulate_ate_data_with_weights(
  n_sample = 10000,
  n_population = 100000,
  p_z_sample = 0.1,
  p_z_population = 0.5,
  beta_a = 1,
  beta_z = 2.5,
  noise_sd = 0.5
)
```

Ok, we have generated both sample and population data.

Next, we verify that the distributions of effect modifiers differ in the sample and in the target population:

```r
# obtain the generated data
sample_data <- data$sample_data
population_data <- data$population_data


# check imbalance
table(sample_data$z_sample) # type 1 is rare
```

```
   0    1
9055  945
```

```r
table(population_data$z_population) # type 1 is common
```

```
    0     1
49916 50084
```

Good, the distributions differ. The simulation is working as intended.

Next, consider the question: "What are the differences in the coefficients that we obtain from the study population at the end of study, as compared with the target population?"

First, we obtain the coefficients for the sample. They are as follows:

```r
# model coefficients sample
model_sample  <-
  glm(y_sample ~ a_sample * z_sample, data = sample_data)

# summary
parameters::model_parameters(model_sample, ci_method = "wald")
```

```
Parameter           | Coefficient |       SE |        95% CI | t(9996) |      p
-------------------------------------------------------------------------------
(Intercept)         |   -6.89e-03 | 7.38e-03 | [-0.02, 0.01] |   -0.93 | 0.350
a sample            |        1.01 |     0.01 | [ 0.99, 1.03] |   95.84 | < .001
z sample            |        2.47 |     0.02 | [ 2.43, 2.52] |  104.09 | < .001
a sample × z sample |        0.51 |     0.03 | [ 0.44, 0.57] |   14.82 | < .001
```

Ok, let's obtain the coefficients for the weighted regression of the sample. Notice that the coefficients are virtually the same:

```r
# model the sample weighted to the population, again note that these coefficients are similar
model_weighted_sample <-
  glm(y_sample ~  a_sample  * z_sample,
      data = sample_data,
      weights = weights)

# summary
summary(parameters::model_parameters(model_weighted_sample, ci_method =
                                     "wald"))
```

```
Parameter           | Coefficient |        95% CI |      p
-----------------------------------------------------------
(Intercept)         |   -6.89e-03 | [-0.03, 0.01] | 0.480
```

```
a sample            |        1.01 | [ 0.98, 1.04] | < .001
z sample            |        2.47 | [ 2.45, 2.50] | < .001
a sample × z sample |        0.51 | [ 0.47, 0.55] | < .001

Model: y_sample ~ a_sample * z_sample (10000 Observations)
Residual standard deviation: 0.494 (df = 9996)
```

We might be tempted to infer that weighting wasn't relevant to the analysis. However, we'll see that such an interpretation would be a mistake.

Next, let us obtain model coefficients for the population. Note again there is no difference – only narrower errors owing to the large sample size.

```
# model coefficients population -- note that these coefficients are very similar.
model_population <-
  glm(y_population ~ a_population * z_population, data = population_data)

parameters::model_parameters(model_population, ci_method = "wald")
```

```
Parameter                   | Coefficient |       SE |        95% CI | t(99996) |      p
-------------------------------------------------------------------------------------------
(Intercept)                 |    2.49e-03 | 3.18e-03 | [ 0.00, 0.01] |     0.78 | 0.434
a population                |        1.00 | 4.49e-03 | [ 0.99, 1.01] |   222.35 | < .001
z population                |        2.50 | 4.49e-03 | [ 2.49, 2.51] |   556.80 | < .001
a population × z population |        0.50 | 6.35e-03 | [ 0.49, 0.51] |    78.80 | < .001
```

Again, there is no difference. That is, we find that all model coefficients are practically equivalent. The different distribution of effect modifiers does not result in different coefficient values for the treatment effect, the effect-modifier "effect," or the interaction of effect modifier and treatment.

Consider why this is the case: in a large sample where the causal effects are invariant – as we have simulated them to be – we will have good replication in the effect modifiers within the sample, so our statistical model can recover the *coefficients* for the population – no problem.

However, **in causal inference, we are interested in obtaining the marginal effect of the treatment**. That is, we seek an estimate for the counterfactual contrast in which everyone in a pre-specified population was subject to one level of treatment compared with a counterfactual condition in which everyone in a population was subject to another level of the same treatment. **When the sample population differs in the distribution of effect modifiers from the target population effect, the marginal effect estimates will typically differ.**

To see this, we use the `stdReg` package to recover marginal effect estimates, comparing (1) the sample ATE, (2) the true oracle ATE for the population, and (3) the weighted sample ATE. We will use the outputs of the same models above. The only difference is that we will calculate marginal effects from these outputs. We will contrast a difference from an intervention in which everyone receives treatment = 0 with one in which everyone receives treatment = 1, however, this choice is arbitrary, and the general lessons apply irrespective of the estimand.

First, consider this ATE for the sample population.

```
# What inference do we draw?  We cannot say the models are unbiased for the marginal effect estimates.
# regression standardisation
library(stdReg) # to obtain marginal effects


# obtain sample ate
fit_std_sample <-
```

```
  stdReg::stdGlm(model_sample, data = sample_data, X = "a_sample")

# summary
summary(fit_std_sample,
        contrast = "difference",
        reference = 0)
```

```
Formula: y_sample ~ a_sample * z_sample
Family: gaussian
Link function: identity
Exposure:  a_sample
Reference level:  a_sample = 0
Contrast:  difference

  Estimate Std. Error lower 0.95 upper 0.95
0     0.00     0.0000       0.00       0.00
1     1.06     0.0101       1.04       1.08
```

The treatment effect is given as a 1.06 unit change in the outcome across the sample population, with a confidence interval from 1.04 to 1.08.

Next, we obtain the true (oracle) treatment effect for the population under the same intervention.

```
## note the population effect is different

#obtain true ate
fit_std_population <-
  stdReg::stdGlm(model_population, data = population_data, X = "a_population")

# summary
summary(fit_std_population,
        contrast = "difference",
        reference = 0)
```

```
Formula: y_population ~ a_population * z_population
Family: gaussian
Link function: identity
Exposure:  a_population
Reference level:  a_population = 0
Contrast:  difference

  Estimate Std. Error lower 0.95 upper 0.95
0     0.00    0.00000       0.00       0.00
1     1.25    0.00327       1.24       1.26
```

Behold, the true treatment effect is a 1.25 unit change in the population, with a confidence bound between 1.24 and 1.26. This is well outside the ATE that we obtain from the sample population!

Next, consider the ATE in the weighted regression, where the sample was weighted to the target population's true distribution of effect modifiers.

```r
## next try weights adjusted ate where we correctly assign population weights to the sample
fit_std_weighted_sample_weights <- stdReg::stdGlm( model_weighted_sample,
    data = sample_data,
    X = "a_sample")

# this gives us the right answer
summary(fit_std_weighted_sample_weights,
    contrast = "difference",
    reference = 0)
```

```
Formula: y_sample ~ a_sample * z_sample
Family: gaussian
Link function: identity
Exposure:  a_sample
Reference level:  a_sample = 0
Contrast:  difference

  Estimate Std. Error lower 0.95 upper 0.95
0     0.00     0.0000       0.00       0.00
1     1.25     0.0172       1.22       1.29
```

```r
# Moral of the story. When we marginalise over the entire sample we need to weight estimates to the target popul
```

Good news, we find that we obtain the population-level causal effect estimate with accurate coverage by weighting the sample to the target population. So with appropriate weights, our results generalise from the sample to the target population.

Lessons

- Regression coefficients do not clarify the problem of sample/target population mismatch – or selection bias as discussed in this manuscript.
- The correct advice to investigators is that they should not rely on regression coefficients when evaluating the biases that arise from sample attrition. This advice applies to both methods that the authors use to investigate threats of bias. That is, to implement this advice, the authors must first take it.
- Generally, observed data are insufficient for assessing threats. Observed data do not clarify structural sources of bias, nor do they clarify effect-modification in the full counterfactual data condition in which all receive the treatment and all do not receive the treatment (at the same level).
- To properly assess bias, one would need access to the counterfactual outcome—what would have happened to the missing participants had they not been lost to follow-up or had they responded. Again, such counterfactual or "full data" are inherently unobservable (Van Der Laan and Rose 2011).
- In simple settings like the one we just simulated, we may address the gap between the sample and target population using methods such as modelling the censoring (e.g., censoring weighting). However, we never know what setting we are in or whether it is simple—such modelling must be handled with care. There is a large and growing epidemiology literature on this topic (see, for example, Li *et al.* (2023)).
- Matters become more complex when there is confounding and selection bias because the problem is not merely one of external validity but also internal validity (i.e. obtaining valid causal effect estimates for the baseline sample). See for example Scharfstein *et al.* (1999); Laan and Robins (2003); Howe *et al.* (2016) (note that Howe *et al.* (2016), "selection" bias is defined as collider stratification bias, a variety of confounding bias and not simple sample and target population mismatch.)

- Matters become more complex still in the presence of treatment-confounder feedback, for which only special methods are needed e.g. Rotnitzky *et al.* (2017) chapters in Laan and Gruber (2012).

Suggestions for revision

Assumptions and validity in specific contexts

- I think the authors must clarify the conditions under which their methodology is applicable. Theoretically, if they wanted to retain the current analysis, they might acknowledge that their approach could hold under unverifiable assumptions of Missing Completely At Random (MCAR) and the absence of interactions. Although such a commentary may be relevant for biological inquiries, such as vaccine efficacy, where these assumptions might be more plausible, it seems rather doubtful in the study of religions.

Utility of regression coefficients

- I recommend critically examining how regression coefficients may mislead in assessing external validity. By illustrating the problem through simulations—along the lines suggested here—the authors could offer a valuable perspective. Such an approach would not only elucidate the methodological pitfalls but also enrich understanding of sample/target population mismatch. For example, the authors could re-examine the assumptions of previous studies.

Address the "Elephant in the Room"

- The discussion should extend to the broader implications of sample/target population mismatch beyond the immediate context of pregnant women in Bristol during the early 1990s. Highlighting this issue encourages clarity about assessing the generalisability of causal effect estimates. To date, the small but growing literature in this area has not been careful on this point.

By addressing these points, a substantially revised manuscript has the potential to make a substantive contribution to the literature on causal inference and its application to the study of religion and culture more generally.

Small Points

1. Abstract accuracy

- Non-random participation is not a unitary phenomenon, but a spectrum of challenges. These include defining or selecting the target population, varying participation rates, nested vs. non-nested trial designs & others. The authors might find the following works helpful: Dahabreh and Hernán (2019) and Dahabreh *et al.* (2021) for randomised trials and Bareinboim and Pearl (2013) for observational studies.

2. Abstract is unclear

- The abstract should promptly and clearly state the study's objective, focusing on the effect of church attendance on depression among the target population. The phrase "despite non-random participation by the exposure and outcome" is unclear because the meanings of these terms have yet to be defined.

3. Clarify the target population throughout

- In abstract and throughout, clearly define the target population for ALSPAC at baseline, for example, as "healthy pregnant women living in Bristol in the early 1990s". Address the generalisation concern explicitly.

4. Revise redundant statements

- For example, "…with attendance at a place of worship associated with continued participation…" This sounds like "with attendance at a place of worship associated with attendance at a place of worship."

5. Simplify terminology

- Consider consistently using a term such as "composite religiosity" instead of the acronym *RSBB* throughout the article for ease of understanding.

6. Clarify methodology

- The introduction to the simulation approach should explicitly address potential concerns about effect modification and how it influences the study's findings.
- If you retain comparative data analysis (I think you should not), explain each step of the analysis clearly.

7. Clarify rationale for variable selection

- The rationale for selecting a limited number of variables from ALSPAC should be clearly explained. Arguably, a richer dataset is needed to assess effect-modification. Moreover, confounding would need to be addressed as well—if internal validity fails, so too will external validity. The authors cannot lose sight of confounding bias if their interest is in external validity. Again, I suggest simulating data in which a mismatch arises even without confounding bias.

9. Directed Acyclic Graph – not persuasive, and probably not needed.

- If you keep the DAG, make sure to avoid cycles (e.g., by indexing nodes by time); reconsider the credibility of the assumptions made, especially concerning effect modification by ethnicity (only through SEP and Marital Status)? Again, problems arise even with unconfoundedness, as demonstrated above, so addressing confounding bias is insufficient to address threats to generalisation from attrition. A simple contribution would set confounding bias to the side.

References

Bareinboim, E, and Pearl, J (2013) A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, **1**(1), 107–134.

Bulbulia, JA (2024) *Margot: MARGinal observational treatment-effects.* doi:10.5281/zenodo.10907724.

Dahabreh, IJ, Haneuse, SJA, Robins, JM, … Hernán, MA (2021) Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology*, **190**(8), 1632–1642.

Dahabreh, IJ, and Hernán, MA (2019) Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, **34**(8), 719–722. doi:10.1007/s10654-019-00533-2.

Howe, CJ, Cole, SR, Lau, B, Napravnik, S, and Eron Jr, JJ (2016) Selection bias due to loss to follow up in cohort studies. *Epidemiology*, **27**(1), 91–97.

Laan, MJ van der, and Gruber, S (2012) Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, **8**(1).

Laan, MJ, and Robins, JM (2003) *Unified methods for censored longitudinal data and causality*, Springer.

Li, W, Miao, W, and Tchetgen Tchetgen, E (2023) Non-parametric inference about mean functionals of non-ignorable non-response data without identifying the joint distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **85**(3), 913–935.

Lüdecke, D, Ben-Shachar, MS, Patil, I, and Makowski, D (2020) Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, **5**(53), 2445. doi:10.21105/joss.02445.

Rotnitzky, A, Robins, J, and Babino, L (2017) On the multiply robust estimation of the mean of the g-functional. *arXiv Preprint arXiv:1705.08582*.

Scharfstein, DO, Rotnitzky, A, and Robins, JM (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94**(448), 1096–1120.

Sjölander, A (2016) Regression standardization with the R package stdReg. *European Journal of Epidemiology*, **31**(6), 563–574. doi:10.1007/s10654-016-0157-3.

Van Der Laan, MJ, and Rose, S (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data*, New York, NY: Springer. Retrieved from https://link.springer.com/10.1007/978-1-4419-9782-1