Introduction to the use of weights for sampling/selection bias

Lorenzo Fabbri Xavier Basagana Martine Vrijheid

2024-03-15

Table of contents

1 Selection bias

From the subjects belonging to the HELIX sub-cohort, only a fraction took part to the follow-up (as part of the ATHLETE project). That is, among the eligible subjects, some are going to be excluded from our analyses since they have e.g., no outcome. Censoring from the analysis those with missing values will most likely introduce **selection bias**.

1.1 A simple example

Suppose exposure `A` is a measure of socio-economic position (SEP), and outcome `Y` is diagnosis of ADHD. The censoring variable `C` is a collider on this pathway, and `L` is a confounder. To estimate the effect of `A` on `Y` we should NOT adjust for `C` (Figure 1).

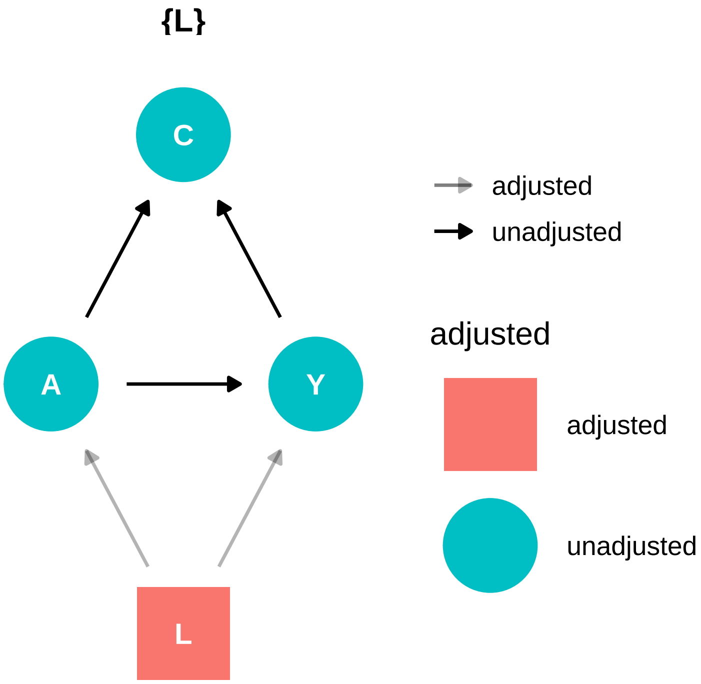

Figure 1: Effect estimation with selection bias

The problem is that in the follow-up, we are implicitly adjusting for `C` as well, thus opening another path. Thus, censoring due to loss to follow-up can introduce selection bias. Generally, we are interested then in estimating an effect if nobody had been censored. If `A` is binary, we are interested in:

$$E\left[Y^{a=1,c=0}\right] - E\left[Y^{a=0,c=0}\right],$$

which is the joint effect of `A` and `C`.

1.2 Some strategies

There are multiple strategies available to reduce the effect of this selection bias on the association between exposure and outcome. In what follows, we assume that the identifiability conditions for the joint treatment (A, C) conditional on L hold.

One such strategy consists in computing **inverse-probability (IP) weights**. In this case, we are interested in estimating the parameters of the marginal structural model (MSM) $E\left[Y^{a,c=0}\right] = \beta_0 + \beta_1 a$, with `a` being the exposure and `c` a indicator variable for censoring. Then, the IP weights are defined as follows:

$$W^{A,C} = W^{A} \times W^{C}, \tag{1}$$

with $W^C = 1/Pr\left[C = 0|L, A\right]$ ($W^C = 0$ for the censored), and W^A being the weights to adjust for confounding ($f(A|L)$). Alternatively, we can compute *stabilized* IP weights:

$$SW^C = \frac{Pr\left[C = 0|A\right]}{Pr\left[C = 0|L, A\right]}.$$

With IP weights, we are effectively modeling the censoring mechanism given the exposure and a set of covariates (e.g., sex and age). Another strategy consists in directly modeling the outcome (standardization). In this case, instead of using weights to adjust for confounding and selection bias, we are directly including theses variables in the outcome model:

$$E\left[Y|A = a, C = 0, L = l\right].$$

Another strategy consists in modeling both the censoring mechanism and the outcome, and *combine* the results together. In this case, we would fit a weighted outcome model, where the weights are as in Equation 1. The same set of covariates used to estimate W^C should be incorporated also in the outcome model.

2 Estimating the balancing weights

In what follows, we will focus on how to estimate these balancing weights, and how to use them when fitting the outcome model. Since weight estimation depends on question-specific confounders and eventually exposures, it makes sense for the researcher to estimate them directly, rather than providing them for *general purposes*.

2.1 Step-by-step procedure

We will make use of the `WeightIt` R package to estimate the IP weights (Greifer 2024b). This package offers a multitude of methods to estimate the balancing weights, from a traditional generalized linear model to data-adaptive methods that do not make use of any parametric specification. The researcher is suggested to try different methods and assess the resulting balance with strategies outlined below.

We will make use of the Lalonde dataset, and we will estimate the effect of `treat` on the continuous outcome `re78`. We will assume that `age`, `educ`, `nodegree`, and `race` are sufficient to remove any confounding bias. We will further create a new (random) variable, `c`, to indicate whether each subject was either censored ($c = 1$) or not ($c = 0$), based on `educ` and `age`. This is a very simplistic scenario to illustrate the basic steps.

```r
type_exposure <- "binary"

dat <- cobalt::lalonde |>
  tibble::as_tibble() |>
  tidylog::mutate(
    censoring_mech = age * educ
  )
dat$c <- with(dat, rbinom(
  n = nrow(dat),
  size = 1,
  prob = (censoring_mech - min(censoring_mech)) /
    (max(censoring_mech) - min(censoring_mech))
))
dat$c <- as.factor(dat$c)
```

We can now proceed to estimate the balancing weights for selection bias, assuming that the censoring mechanism solely involves the variables `educ`, `age`, and `race`. We will fit a simple generalized linear model, although it is recommended to test also other methods that do not make such strong parametric assumptions (e.g., the `energy` method). Some of the estimated weights might be considered *extreme* (you can check this with the `summary` function applied to the returning object, which provides other useful information like the effective sample size before and after weighting). We therefore might want to trim/winsorize them.

```r
est_sel_weights <- WeightIt::weightit(
  c ~ educ + age + race,
  data = dat,
  method = "glm"
)
summary(est_sel_weights)
```

```
                 Summary of weights

- Weight ranges:

     Min                                  Max
0 1.0709 |---------------|             4.8724
1 1.1732 |---------------------------| 7.3521

- Units with the 5 most extreme weights by group:

      405    276    310    296    249
 0 3.3047 3.5212 3.6902 4.2847 4.8724
      605    398     61     47    543
 1 6.3039 6.6955 6.9197 6.9197 7.3521

- Weight statistics:

  Coef of Var   MAD Entropy # Zeros
0       0.300 0.203   0.038       0
1       0.483 0.372   0.105       0

- Effective Sample Sizes:

                0      1
Unweighted 379.   235.
Weighted   347.81 190.72
```

```r
est_sel_weights_trim <- WeightIt::trim(
  est_sel_weights,
  at = 0.9,
  lower = TRUE
)
summary(est_sel_weights_trim)
```

```
                 Summary of weights

- Weight ranges:

     Min                                 Max
0 1.2187 |---------------------------| 3.3054
1 1.2187 |---------------------------| 3.3054

- Units with the 5 most extreme weights by group:

      405    310    296    276    249
 0 3.3047 3.3054 3.3054 3.3054 3.3054
       61     51     47     45     18
 1 3.3054 3.3054 3.3054 3.3054 3.3054

- Weight statistics:

  Coef of Var   MAD Entropy # Zeros
0       0.270 0.194   0.032       0
1       0.316 0.277   0.051       0

- Effective Sample Sizes:

                 0      1
Unweighted 379.   235.  
Weighted   353.31 213.72
```

In Figure 2 you can see the distribution of the propensity scores for both *treatment* levels, in the unadjusted and adjusted samples.

```r
cobalt::bal.plot(
  est_sel_weights_trim,
  var.name = "prop.score",
  which = "both",
  type = "histogram",
  mirror = TRUE
)
```

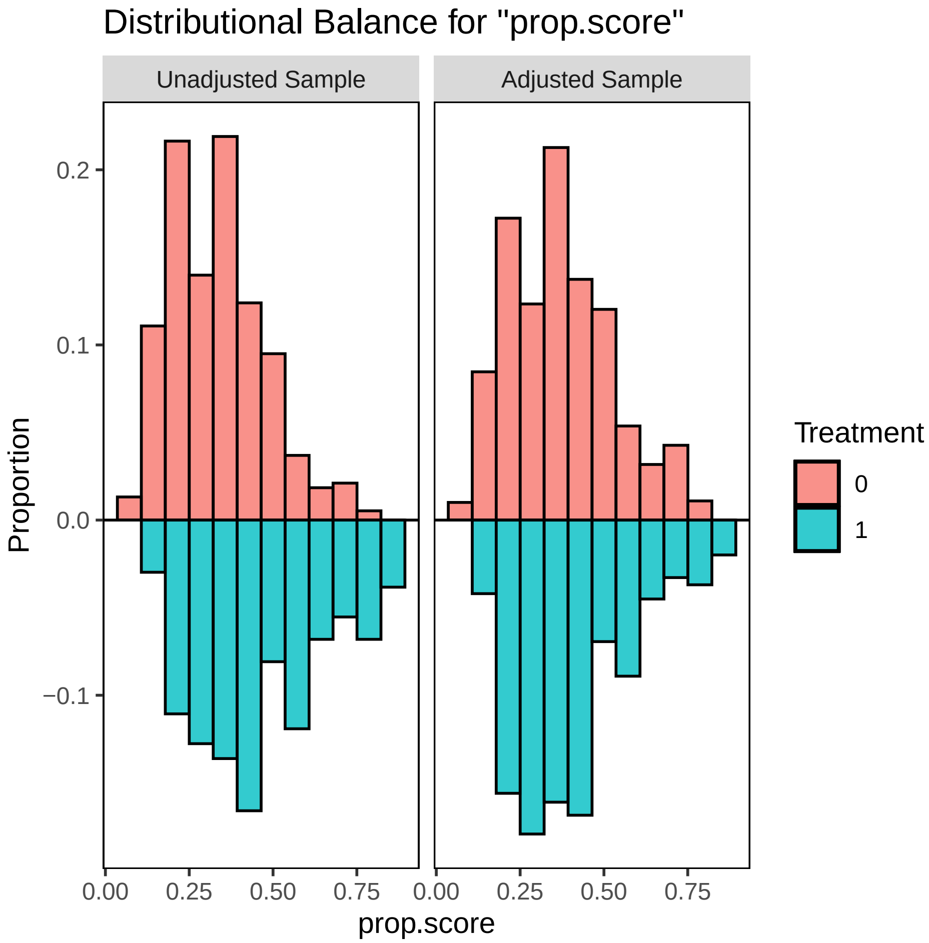

Figure 2: Propensity scores

We can visually check whether the estimated weights are effectively capable of reducing the effect of the chosen covariates on the censoring mechanism using a so-called *Love plot* (Figure 3). Ideally, we would like to see that the adjusted values are as close as possible to zero. Deviations from this should be solved for instance by choosing another fitting method and/or by modifying the trimming value.

```
cobalt::love.plot(
  est_sel_weights_trim,
  stats = ifelse(
    type_exposure == "continuous",
    c("correlations"),
    c("mean.diffs")
  ),
  binary = "std",
  abs = TRUE,
  var.order = "unadjusted",
  thresholds = c(cor = 0.1, m = 0.1),
  line = TRUE,
)
```

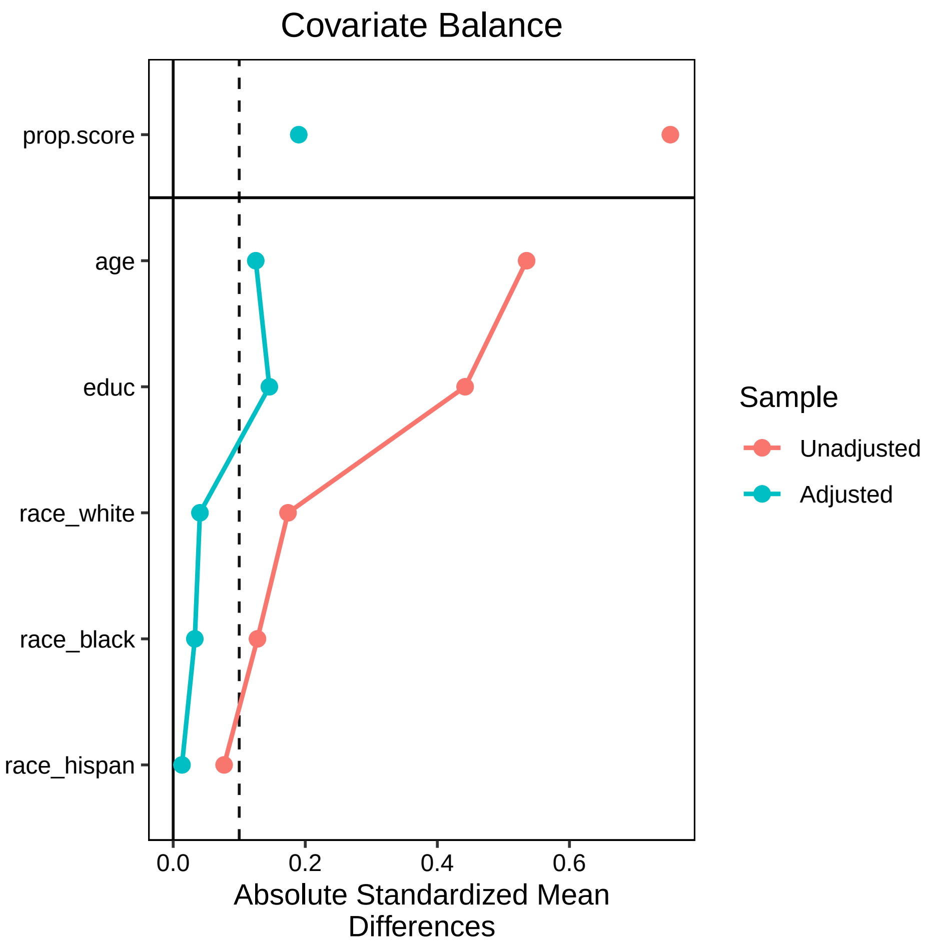

Figure 3: Love plot

Once we are satisfied with the estimated balancing weights, we can proceed to fit our (weighted) outcome model.

```r
fit <- glm(
  formula = re78 ~ treat + age + educ + nodegree + race,
  weights = est_sel_weights_trim$weights,
  data = dat
)
```

We can compare the estimated coefficient of `treat` with that of a model without weights (Table 1).

```r
fit_null <- glm(
  formula = re78 ~ treat + age + educ + nodegree + race,
  weights = NULL,
  data = dat
)
```

Table 1: Comparison of estimated coefficients

	No weights			With weights		
Characteristic	**Beta**	**95% CI**[1]	**p-value**	**Beta**	**95% CI**[1]	**p-value**
treat	829	-759, 2,418	0.3	914	-716, 2,544	0.3

[1] CI = Confidence Interval

2.2 The `SelectionWeights` R package

TBD.

3 Suggestions

3.1 Selection of covariates

We strongly advise against the practice of selecting covariates to estimate the weights based on *p*-values of the associations between covariates and censoring. Instead, these variables should be selected based on *a priori* knowledge. Covariates that might be involved in the censoring mechanism include: cohort, sex and age of the subject, socioeconomic factors, season of visit, health status of the subject (e.g., presence of a debilitating health condition). It is also important to note that not all missing values are created equal: the reason why a individual did not take part in the follow-up might also be relevant (e.g., not wanting to take part or not being able to contact them).

We would like to further provide some suggestions based on previous experience.

- There might be significant differences in terms of censoring mechanism between cohorts. It might be useful to estimate the balancing weights for each cohort separately.
- Consider computing standard errors using a robust method (see vignette).
- It might be useful to provide further information about the balancing. More tools are available in the `cobalt` R package (Greifer 2024a).
- It might be useful to compare the results (e.g., predicted outcome or exposure coefficient) with and without IP weights, to check whether something went wrong during the estimation process. This information can be put in the supplementary material.

References

Alten, Sjoerd van, Benjamin W Domingue, Titus Galama, and Andries T Marees. 2022. "Reweighting the UK Biobank to Reflect Its Underlying Sampling Population Substantially Reduces Pervasive Selection Bias Due to Volunteering." *medRxiv*, 2022–05.

Greifer, Noah. 2024a. *Cobalt: Covariate Balance Tables and Plots*. https://ngreifer.github.io/cobalt/.

———. 2024b. *WeightIt: Weighting for Covariate Balance in Observational Studies*. https://ngreifer.github.io/WeightIt/.

Hernán, Miguel A, and James M Robins. 2010. "Causal Inference." CRC Boca Raton, FL.