**Principal Surrogate Evaluation with pseval**

**Basics**

*pseval* is designed to analyze data from a randomized clinical trial in order to assess the surrogate value of a post-randomization measurement. Start by describing the study design, including augmentations.

\[
\text{pl} \leftarrow \text{psdesign} \left( \text{data} = \text{data}, Z = Z, Y = Y.\text{obs}, S = S.\text{obs}, \text{BIP} = \text{BIP}, \text{CPV} = \text{CPV} \right)
\]

The counterfactual surrogate \(S.1\) is missing for many subjects, thus we need to define a model to integrate over the missing values.

\[
\text{pl} \leftarrow \text{pl} + \text{integrate\_parametric}(S.1 \sim \text{BIP})
\]

The risk model describes the relationship between the outcome \(Y\), the surrogate \(S.1\), and the treatment \(Z\). Use the risk model that is most appropriate for your outcome type, binary, count, or time-to-event.

\[
\text{pl} \leftarrow \text{pl} + \text{risk\_binary}(Y \sim S.1 \ast Z, D=500, \text{risk.logit})
\]

**Study Design** specification and mapping

*psdesign* controls the dataset that is being used, and how to map variables to their roles in the analysis. The “keys” to the left of “=“ map to variables in data

- **DataFrame**
  \[
  a \leftarrow \text{psdesign} \left( \text{data} = \text{data}, Z = Z, Y = Y.\text{obs}, S = S.\text{obs}, \text{BIP} = \text{BIP} \right)
  \]

- **Treatment**
  \[
  b \leftarrow \text{psdesign} \left( \text{data} = \text{data}, Z = Z, Y = \text{Surv(time.\text{obs}, \text{event.\text{obs}})}, \tau = .25, S = S.\text{obs}, \text{BIP} = \text{W}, \text{CPV} = \text{CPV}, \text{BSM} = \text{V1}, \text{weights} = \text{p}, \text{covariate} = X \right)
  \]

**Integration** over the missing counterfactuals

inegrate\_* functions control how the missing counterfactual variables are handled

- **Parametric**: Assumes normal distribution conditional on a BIP + other variables
  \[
  a + \text{integrate\_parametric}(S.1 \sim \text{BIP})
  \]

- **Semiparametric**: Assumes location and scale vary as functions of BIP + other variables, no assumption about distribution of \(S\)
  \[
  a + \text{integrate\_semiparametric(}
  \text{formula.location} = S.1 \sim \text{BIP},
  \text{formula.scale} = S.1 \sim 1
  \)

- **Nonparametric**: Totally empirical, requires categorical \(S\) and \(W\)
  \[
  a + \text{integrate\_nonparametric}(S.1 \sim \text{BIP})
  \]

**Risk Model** distribution of the outcome

risk\_* functions define the assumed relationship between \(Y, S.1,\) and \(Z\). The default model is \(Y \sim S.1 \ast Z\)

- **Binary outcome**
  \[
  a + \text{risk\_binary}(\text{risk} = \text{risk.logit})
  \]

- **Time to event outcome**
  \[
  a + \text{risk\_exponential()}
  \]

- **Count outcome**
  \[
  a + \text{risk\_poisson()}
  \]

**Options**

- **Flexible Spline**
  \[
  a + \text{risk\_binary}(Y \sim \text{bs}(S.1, df = 2) \ast Z)
  \]

**Estimation** post-estimation and plotting

est \(\leftarrow a + \text{ps\_estimate(method = “BFGS”)}\)

\[
\text{boot} \leftarrow \text{est} + \text{ps\_bootstrap}(\text{n.\text{boots} = 50}, \text{start} = \text{binary.\text{est}\$estimates\$par})
\]

**Post estimation**

summary(boot)

\[
\text{calc\_risk}(\text{boot}, \text{contrast} = “VE”)
\]

\[
\text{calc\_STG}(\text{boot}) \# \text{total gain statistic}
\]

plot(boot, contrast = “VE”)

\[
\text{plot}(\text{boot}, \text{contrast} = “logRR”)
\]

\[
\text{plot}(\text{boot}, \text{contrast} = “RD”), \text{CI.type} = “pointwise”)
\]

\[
\text{calc\_risk}(\text{boot}, \text{contrast} = \text{function(R0, R1) 1 - R1/R0})
\]

[See ?optim for options]