Introduction

Pseudo-observations provide an approach for investigating how a parameter θ, expressible as the expectation of the outcome variable Y, depends on a set of covariates X. Let (Y_i, X_i) denote IID observations. Consider the parameter defined by:

$$\theta = \mathbb{E}_Y\{g(Y_i)\},$$

where g is a known function. For each observation i, define the conditional value:

$$\theta_i = \theta(X_i) = \mathbb{E}_{Y|X}\{g(Y_i)|X_i\},$$

and note that $\theta = \mathbb{E}_X(\theta_i)$. The ith **pseudo-observation** is defined as:

$$\hat{\theta}_i = \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}^{(-i)}) = n\hat{\theta} - (n-1)\hat{\theta}^{(-i)}. \tag{1.1}$$

where $\hat{\theta}$ is the full-sample estimate of θ, and $\hat{\theta}^{(-i)}$ is the **jackknife** estimate obtained by omitting observation i.

Pseudo-observations are particularly useful in the survival setting, where the outcome data are subject to censoring. Quantities amenable to modeling using pseudo-observations include the survival probability, the restricted mean survival time, and transition probabilities in multi-state models. The pseudo-observations are generated in a way that accounts for censoring. However, once generated, the pseudo-observations can be directly modeled within the standard generalized estimating equation framework; no further specialization to account for censoring is required.

Pseudo-observation calculation

The pseudo-observation $\hat{\theta}_i$ for subject i is a measure of how much influence subject i has on the full-sample estimate $\hat{\theta}$. Rearranging (1.1) gives:

$$\hat{\theta}_i - \hat{\theta} = (n-1)\big\{\hat{\theta} - \hat{\theta}^{(-i)}\big\} = \frac{T\{\mathbb{F}_n + \epsilon_n(\delta_i - \mathbb{F}_n)\} - T(\mathbb{F}_n)}{\epsilon_n},$$

where $\hat{\theta} = T(\mathbb{F}_n)$ expresses $\theta = T(F)$ as a statistical functional, δ_i is a point-mass on observation i, and $\epsilon_n = -(n-1)^{-1}$. Recall that the **influence function** for observation i is defined by:

$$\psi_i = \lim_{\epsilon \to 0} \frac{T\{F + \epsilon(\delta_i - F)\} - T(F)}{\epsilon}$$

Thus, $\hat{\theta}_i - \hat{\theta}$ is a finite-sample approximation to ψ_i, which improves for $n \to \infty$. Moreover, if ψ is known, than the pseudo-observations can be calculated (approximately) without jackknifing:

$$\hat{\theta}_i \approx \hat{\theta} + \psi_i$$

2.1 Kaplan-Meier

Consider right-censored survival data (U_i, δ_i), $U_i = \min(T_i, C_i)$, $\delta_i = \mathbb{I}(T_i \le C_i)$. The influence function expansion for the Kaplan-Meier estimator $\hat{S}$ is:

$$\frac{\sqrt{n}\{\hat{S}(t) - S(t)\}}{-S(t)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \frac{dM_i(u)}{n^{-1}\sum_{j=1}^{n} Y_j(u)} + o_p(1)$$

where:

$$Y_i(u) = \mathbb{I}(U_i \ge u)$$

is the at-risk process,

$$N_i(u) = \mathbb{I}(U_i \le u, \delta_i = 1)$$

is the counting process,

$$dM_i(u) = dN_i(u) - Y_i(u)dA(u)$$

is the counting process martingale, and $A(u)$ is the cumulative hazard. The influence function for subject i with respect to the KM curve is:

$$\psi_i(t) = -S(t) \int_0^t \frac{dM_i(u)}{n^{-1}Y(u)},$$

where $Y(u) = \sum_{j=1}^{n} Y_j(u)$ is the total number at risk. Thus, for $\theta(t) = \mathbb{P}(T > t)$, the pseudo-observation can be approximated as:

$$\hat{\theta}_i(t) \approx \hat{S}(t) - \hat{S}(t) \int_0^t \frac{d\hat{M}_i(u)}{n^{-1}Y(u)}.$$

2.2 Restricted mean survival

The restricted mean survival time is:

$$R(\tau) = \mathbb{E}\{T \wedge \tau\} = \int_0^{\tau} S(u)du.$$

Defining $\theta(\tau) = R(\tau)$ and building directly upon the Kaplan-Meier pseudo-observations, those for the RMST can be approximated as:

$$\hat{\theta}_i(\tau) = \int_0^{\tau} \hat{S}(t)dt - \int_0^{\tau} \hat{S}(t) \int_0^t \frac{d\hat{M}_i(u)}{n^{-1}Y(u)}.$$

Alternatively, using integration by parts, the influence function for the RMST can be expressed as:

$$\psi_i(t) = -\int_0^{\tau} \frac{\mu_{\tau}(u)dM_i(u)}{n^{-1}Y(u)},$$

where:

$$\mu_\tau(u) = \int_u^\tau S(t)dt.$$

Thus, the RMST pseudo-observations are alternatively given by:

$$\hat{\theta}_i(\tau) = \int_0^\tau \hat{S}(t)dt - \int_0^\tau \frac{\hat{\mu}_\tau(t)d\hat{M}_i(t)}{n^{-1}Y(t)}.$$

2.3 Cumulative incidence function

Suppose $\delta_i \in \{0, 1, \cdots, J\}$ can assume $(J+1)$ possible values, where $\delta_i = 0$ corresponds to censoring and $\delta_i = j \geq 1$ corresponds to the jth event of interest. The cumulative incidence of the jth event is:

$$F_j(t) = \mathbb{P}(T_i \leq t, \delta_i = j),$$

which is estimated by:

$$\hat{F}_j(t) = \int_0^t \exp\left\{-\sum_{j=1}^J \frac{dN_j(u)}{Y(u)}\right\} \frac{dN_j(u)}{Y(u)}.$$

The influence function for observation i on $\hat{F}_j$ is:

$$\psi_{ji}(t) = -F_j(t)\int_0^t \frac{dM_i(u)}{n^{-1}Y(u)} + \int_0^t \frac{F_j(u)dM_i(u)}{n^{-1}Y(u)} + \int_0^t \frac{S(u)dM_{ji}(u)}{n^{-1}Y(u)},$$

where:

$$dM_{ji}(u) = dN_{ji}(u) - \int_0^u Y_i(t)dA_j(t)$$

is the cause-specific martingale, and:

$$dM_i(u) = \sum_{j=1}^J dM_{ij}(u) = dN_i(u) - \int_0^u Y_i(t)dA(t).$$

Additionally,

$$N(t) = \sum_{j=1}^J N_j(t), \qquad A(t) = \int_0^t \frac{dN(u)}{Y(u)}, \qquad S(t) = \prod_0^t \left\{1 - \frac{dN(u)}{Y(u)}\right\}.$$

Finally, the pseudo-observations can be approximated as:

$$\hat{\theta}_{ji}(t) = \hat{F}_j(t) + \psi_{ji}(t).$$

Modeling pseudo-observations

Having generated pseudo-observations $(\hat{\theta}_i)$, the dependence of θ on covariates X is investigated by specifying a generalized linear model for θ_i:

$$\theta_i = \theta(X_i) = h(\beta' X_i), \tag{3.1}$$

Model (3.1) is estimated within the generalized estimating equation (GEE) framework. Specifically, define the estimating equations:

$$\mathcal{U}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_i = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1}\{\hat{\theta}_i - h(\beta' X_i)\}, \tag{3.2}$$

where $\boldsymbol{D}_i$ is the Jacobian:

$$\boldsymbol{D}_i = \frac{\partial h(\beta' X_i)}{\partial \boldsymbol{\beta'}} = \frac{\partial h(\eta_i)}{\partial \eta_i} X_i',$$

where $\eta_i = \beta' X_i$, and $\boldsymbol{V}_i$ is a working covariance structure, which need to be correctly specified. The estimate $\hat{\beta}$ is obtained by solving $\mathcal{U}(\beta) \overset{\text{Set}}{=} 0$. The variance is obtained using a sandwich estimator:

$$\mathbb{V}(\hat{\beta}) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}, \qquad \boldsymbol{A} = \sum_{i=1}^{n} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i, \qquad \boldsymbol{B} = \sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{U}_i'.$$

The estimates of $\hat{\beta}$ obtained by solving (3.2) are consistent and asymptotically normal:

$$\hat{\beta} - \beta \mathrel{\dot{\sim}} N\{0, \mathbb{V}(\hat{\beta})\}.$$

3.1 Multiple timepoints

For parameters $\theta(t)$ that are time-dependent, pseudo-observations can be generated across a grid of timepoints $(\tau_1, \cdots, \tau_M)$, for instance quantiles of the marginal distribution of T (as estimated by Kaplan-Meier). The GEE framework in (3.2) readily accommodates vector-valued pseudo-observations:

$$\hat{\boldsymbol{\theta}}_i = \begin{pmatrix} \hat{\theta}(\tau_1) \\ \vdots \\ \hat{\theta}(\tau_M) \end{pmatrix}.$$

Modeling multiple timepoints is expected to improve precision when components of β are shared across timepoints. Note that in the multiple timepoint setting, X_i should include time or some function thereof (e.g. a spline basis [1]) to allow θ_i to depend on t. Modeling 10 and not more than 20 timepoints has been recommended [2, 3, 1].

3.2 Working covariance matrix

Klein and Andersen [3] make three suggestions for possible working covariance matrices. A simple but likely inefficient choice for the working covariance matrix is identity $\boldsymbol{V}_i = \boldsymbol{I}$. A more-sophisticated approach is to consider the variance of the estimand in the absence of censoring. For example, consider estimating the survival function:

$$S(t) = \mathbb{P}(T > t) = \mathbb{E}_T\{\mathbb{I}(T > t)\}$$

at multiple time-points $(\tau_1, \cdots, \tau_M)$. In the absence of censoring:

$$\hat{S}(\tau_m) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(T_i > \tau_m).$$

The covariance between any two time points $\tau_{m_1} < \tau_{m_2}$ is:

$$\begin{aligned}
\mathbb{C}\{\hat{S}(\tau_{m_1}), \hat{S}(\tau_{m_2})\} &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{C}\{\mathbb{I}(T_i > \tau_1), \mathbb{I}(T_i > \tau_2)\} \\
&= \frac{1}{n}\Big[\mathbb{E}\{\mathbb{I}(T_i > \tau_1)\mathbb{I}(T_i > \tau_2)\} - \mathbb{E}\{\mathbb{I}(T_i > \tau_1)\}\mathbb{E}\{\mathbb{I}(T_i > \tau_2)\}\Big] \\
&= \frac{1}{n}\{S(\tau_2) - S(\tau_1)S(\tau_2)\} = \frac{1}{n}S(\tau_2)\{1 - S(\tau_1)\}.
\end{aligned}$$

This suggests a working covariance structure of the form:

$$V_{i,ab} = S(\tau_b|X_i)\{1 - S(\tau_a|X_i)\}, \quad a \le b.$$

The $S(t|X_i) = g(\beta' X_i)$ appearing in the working covariance structure may either be updated iteratively during model fitting, or approximated by:

$$\hat{S}(t|X_i) = g(\hat{\beta}_0' X_i),$$

where $\hat{\beta}_0$ is the estimate of β obtained using the working independence model $\boldsymbol{V}_i = \boldsymbol{I}$. A final candidate for $\boldsymbol{V}_i$ is the empirical correlation matrix:

$$\boldsymbol{V}_{i,ab} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_{ia} - \bar{\theta}_a)(\hat{\theta}_{ib} - \bar{\theta}_b), \qquad \bar{\theta}_a = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{ia}.$$

Like the working independence model, the empirical correlation model is not subject-specific.

References

[1] Ambrogi, F and Iacobelli, S and Andersen, PK. "Analyzing differences between restricted mean survival time curves using pseudo-values". In: *BMC Medical Research Methodology* 22.71 (2022). DOI: https://doi.org/10.1186/s12874-022-01559-z.

[2] Andersen, PK and Klein, JP and Rosthoj, S. "Generalised Linear Models for Correlated Pseudo-Observations, with Applications to Multi-State Models". In: *Biometrika* 90.1 (2003), pp. 15–27. URL: https://www.jstor.org/stable/30042016.

[3] Klein, JP and Andersen, PK. "Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function". In: *Biometrics* 61.1 (2006), pp. 223–229. URL: https://www.jstor.org/stable/3695666.