5 Project 4: Linearly Combing Risk Markers to Maximize Standardized Net Benefit

Objective: For this project, we apply the decision-analytic framework to risk model development. We consider a setting where an intervention exits and those with high risk scores are prescribed the intervention. We propose a method for constructing linear combinations of risk markers, $\theta^T \mathbf{X}$ that maximizes a measure of clinical utility, standardized net benefit

5.1 Methods

We propose finding linear combinations of risk markers by directly optimizing equation (8). sNB is a ranked based, and scalings of parameter θ can lead to equivalent decision boundaries. Therefore, we include the constraint $||\theta||^2 = 1$ (Meisner, Fong, others). Given these considerations estimate (θ, t) by solving the optimization problem

$$
\begin{aligned}
(\hat{\theta}, \hat{t}) &= \underset{\theta \in \mathbb{R}^p, t \in \mathbb{R}}{\arg\max} \ \widetilde{sNB}(\theta^T X, t) & (6)\\
&\text{subject to} \ \ ||\theta||^2 = 1, & (7)
\end{aligned}
$$

where $\widetilde{sNB}(\theta^T X, t)$ is a smooth approximation of the empirical estimate of sNB. In particular,

$$
\widetilde{sNB}(\theta, t) = \frac{1}{n_D} \sum_{i=1}^{n_D} \Phi \left(\frac{\theta^T X_i - t}{s_n} \right) - \omega \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n_{\bar{D}}} \Phi \left(\frac{\theta^T X_i - t}{s_n} \right), \tag{8}
$$

where s_n is a tuning parameter that controls the approximation to an indicator function, and $\Phi(x)$ is the standard normal cumulative distribution function. As s_n approaches 0 $\phi(x/s_n)$ more closely approximates $\mathbf{1}[x > 0]$.

This problem can be solved using gradient based measures via the `Rsolpn` software in `R`. Initial values, (θ_0, t_0) are required for software implementation. Fong et al. (2011) and Meisner et al (2017) proposed using normalized estimates of coefficients estimated from robust logistic regression as θ_0. Robust logistic regression can be useful in cases where there are outliers or influential observations that may bias the results of logistic regression. In simulations studies we found, that in some instances standard logistic regression produces linear combinations with higher sNB than robust, and vice versa. For this reason, we propose that sNB be estimated for both the linear combinations obtained for standard and robust logistic regression, and the higher of the two be used to obtain initial values. Note that in order to estimate sNB under standard or robust logistic regression a risk threshold must be provided. To find such a threshold, we solve the one-dimensional optimization problem

$$
\begin{aligned}
t_\beta &= \underset{t \in [0,1]}{\arg\max} \ \widetilde{sNB}(t | \text{expit}(\beta^T) \mathbf{X}) & (9)\\
&= \underset{t \in [0,1]}{\arg\max} \ \frac{1}{n_D} \sum_{i=1}^{n_D} \Phi \left(\frac{\text{expit}(\beta^T X_i) - t}{s_n} \right) - \omega \frac{1}{n_{\bar{D}}} \sum_{i=1}^{n_{\bar{D}}} \Phi \left(\frac{\text{expit}(\beta^T X_i) - t}{s_n} \right). & (10)
\end{aligned}
$$

Once the θ_0 is found, then it is left to find t_0. Similar to solving the one-dimensional optimization problem in (10), the t_0 can be found by finding the t that maximizing $\widetilde{sNB}(t | \theta_0^T \mathbf{X})$.