Predictive Modeling for Heterogeneous Treatment Effects in Medicine

Max Welz
welz@ese.eur.nl

Econometric Institute
Erasmus School of Economics

September 1, 2021

1 Introduction and Notation

1.1 Introduction

The identification of heterogeneity in the effectiveness in medical trials has been traditionally done by *one-variable-at-a-time-analyses*, where heterogeneity is evaluated along one single variable. For a number of statistical reasons, such analyses are insufficient for a reliable identification of treatment effect heterogeneity. Therefore, Kent et al. (2020) recommend to use predictive modeling approaches as well as machine learning techniques for this task. In this document, we introduce the proposed techniques.

1.2 Notation

For the remainder of this paper, let $\{(X_i, Y_i, W_i)\}_{i=1}^n$ be a random sample with the following characteristics. For $i = 1, \dots, n$, we assume that X_i is a p-dimensional random vector of explanatory variables, Y_i is a binary response variable, and W_i is a binary treatment assignment variable. We say that individual i *has the event* if $Y_i = 1$ and that i has been treated if $W_i = 1$. We are interested in the causal effect of W_i on Y_i and potential heterogeneity therein. Using the well-known Rubin causal model, which defined the potential outcomes be defined as $Y_i(1)$ and $Y_i(0)$, denoting the outcome if individual i is treated and if it is not treated, respectively. Recall that we only ever observe one of the potential outcomes for a given i, that is, $Y_i = Y_i(W_i)$. The conditional average treatment effect of individual i at an arbitrary covariate realization $X_i = x$ is formally defined as

$$x \mapsto \theta(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]. \tag{1}$$

We are interested in estimating the treatment effect $x \mapsto \theta(x)$ on at least a group level, based on the observations $\{(X_i, Y_i, W_i)\}_{i=1}^n$. We frequently call $x \mapsto \theta(x)$ the

absolute benefit (of the treatment). Another quantity of interest is

$$x \mapsto \theta^{rel}(x) = \frac{\mathbb{E}[Y_i(1)|X_i = x]}{\mathbb{E}[Y_i(0)|X_i = x]}, \tag{2}$$

which is called the *relative benefit* (of the treatment). At $X_i = x_i$, we often write $\theta_i = \theta(x_i)$ and $\theta_i^{rel} = \theta^{rel}(x_i)$.

2 One-Variable-at-a-Time Analyses

One-variable-at-a-time analyses are typically by means of a series of exact rate ratio tests. Suppose the n samples at hand can be partitioned into two disjoint groups, $\mathcal{G}_0$ and $\mathcal{G}_1$, and these groups are not necessarily collectively exhaustive. The partitioning is usually done according to treatment assignment along some characteristics of a single variable in X_i, hence the name *one-variable-at-a-time analysis*. For example, $\mathcal{G}_0$ and $\mathcal{G}_1$ may contain the younger-than-60-years-old individuals in the control group and treatment group, respectively.

The random variables

$$P_0 = |\{i \in \mathcal{G}_0 : Y_i = 1\}| \quad \text{and} \quad P_1 = |\{i \in \mathcal{G}_1 : Y_i = 1\}|$$

count the number of events in each group. We furthermore assume that we observe the time at risk of of each individual i, denoted T_i. Let

$$N_0 = \sum_{\{i:i\in\mathcal{G}_0\}} T_i \quad \text{and} \quad N_1 = \sum_{\{i:i\in\mathcal{G}_1\}} T_i$$

denote the cumulative time at risk within each group. We assume that the random variables P_0, P_1 that measure number of events in each group are Poisson-distributed as

$$P_0 \sim \mathsf{Poisson}(N_0\lambda_0) \quad \text{and} \quad P_1 \sim \mathsf{Poisson}(N_1\lambda_1),$$

where the parameters $\lambda_0, \lambda_1 > 0$ are unknown. This assumption implies that we can interpret the cumulative number of life years in each group, N_0 and N_1, as the total time spent in the Poisson processes P_0 and P_1, respectively. We are interested in the *rate ratio* θ, defined by

$$\theta = \frac{\lambda_0}{\lambda_1}.$$

Typically, the uniformly most powerful (UMP) test (e.g. p. 152 in Lehmann and Romano, 1986) is used for testing the rate ratio θ. See the vignette of the `rateratio.test` package on the CRAN for details on this test. If we reject the null hypothesis $H_0 : \theta \neq 1$, we have found evidence that there seem to be systematic differences in the occurrence of events between the two groups. For instance, if $Y_i = 1$ denotes that individual i has died and T_i measures the number life years of i, this test can be used to test for systematic differences in mortality rates between the treatment and control group. If a significant difference is found, this difference is due to the treatment intervention (that is at least the idea).

3 Predictive Models

In the following, we introduce the techniques proposed in Kent et al. (2020). This section is partially based on Rekkas et al. (2019). Note that neither Kent et al. (2020) nor Rekkas et al. (2019) propose their methods in an unambiguous way. Hence, this section is an attempt at a mathematically rigorous and statistically sound definition of predictive models. Predictive models can be broken down in *risk models* and *effect models*. The goal of any predictive model is to estimate the absolute and relative benefit in equations (1) and (2), respectively.

3.1 Risk Models

The idea behind risk models is to separately estimate the effect of X_i on Y_i and the effect of W_i on Y_i. In doing so, one separates the explanatory power for Y_i into a part that is due to X_i and a part that is due to W_i. This gives rise to a two-stage estimation procedure, which is known as risk modeling.

3.1.1 Stage 1: Baseline Risk

In stage one, we fit a linear logistic model of X_i to Y_i, which assumes the linear identity

$$\ln\left(\frac{\mathbb{P}[Y_i = 1|X_i = x_i]}{1 - \mathbb{P}[Y_i = 1|X_i = x_i]}\right) = \beta_0 + x_i^\top \beta, \tag{3}$$

for all $i = 1, \ldots, n$, where $(\beta_0, \beta) = (\beta_0, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$ is some fixed but unknown vector of coefficients. Fitting this model by means of cross-validated regularized logistic regression (see Appendix A for details) yields estimates $(\widehat{\beta}_0, \widehat{\beta})$. Due to the regularization penalty in the corresponding optimization problem, the estimate will be sparse; the idea behind this is to separate more relevant predictors from less relevant ones. The estimates $(\widehat{\beta}_0, \widehat{\beta})$ can then be used to compute an estimate of the *baseline risk*, $\widehat{\mathbb{P}}[Y_i = 1|X_i = x_i]$, as well as an estimate of the *linear index*, $\widehat{\eta}_i$, which corresponds to the estimated right-hand side of (3) at $X_i = x_i$:

$$\widehat{\eta}_i = \widehat{\eta}_i(x_i) = \widehat{\beta}_0 + x_i^\top \widehat{\beta}.$$

We retain the linear indices $\widehat{\eta}_i$ for the second stage.

3.1.2 Stage 2: Final Risk

In stage two, we fit another linear logistic model, this time of W_i and the interaction $W_i\widehat{\eta}_i$ to Y_i, where $\widehat{\eta}_i$ is the linear predictor from Stage 1. We moreover assume an individual-specific offset of the value $\widehat{\eta}_i$. Formally, the assumed model satisfies

$$\ln\left(\frac{\mathbb{P}[Y_i = 1|W_i = w_i, \eta_i = \widehat{\eta}_i]}{1 - \mathbb{P}[Y_i = 1|W_i = w_i, \eta_i = \widehat{\eta}_i]}\right) = \alpha_0 + \alpha_1 w_i + \alpha_2 w_i \widehat{\eta}_i + \widehat{\eta}_i \tag{4}$$

for all $i = 1, \ldots, n$, where $\alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$ are coefficients. [There seems to be an inconsistency in the definition of Stage 2 between Rekkas et al. (2019) and Kent

et al. (2020), We've reached out to David Kent on which definition is intended.] We fit this model by means of non-regularized logistic regression. The obtained estimates of the fitted model, $(\widehat{\alpha}_0, \widehat{\alpha}_1, \widehat{\alpha}_2)$, can then be used to calculate the *risk prediction function* for $w \in \{0, 1\}$:

$$\begin{aligned} \mathsf{risk}_i(w) &:= \widehat{\mathbb{P}}[Y_i = 1 | W_i = w, \eta_i = \widehat{\eta}_i] \\ &= \Big(1 + \exp\big(-\widehat{\alpha}_0 - \widehat{\alpha}_1 w - \widehat{\alpha}_2 w \widehat{\eta}_i - \widehat{\eta}_i\big)\Big)^{-1} \\ &= F_{logistic}\Big(\widehat{\alpha}_0 + \widehat{\alpha}_1 w + \widehat{\alpha}_2 w \widehat{\eta}_i + \widehat{\eta}_i\Big), \end{aligned} \tag{5}$$

where $F_{logistic}$ is the distribution function of the logistic distribution with location zero and scale one.

3.2 Effect Models

Effect models attempt to model treatment effect heterogeneity explicitly by adding interaction terms. The idea is that (regularized) logistic regression shrinks the coefficients of interaction terms (and other variables) with low explanatory variables to zero, resulting in a parsimonious model for treatment effect heterogeneity.

However, if one wants to perform inference in a model in which variable selection took place, one needs to address the additional uncertainty stemming from the selection process. This issue is addressed in Wasserman and Roeder (2009). We correspondingly adapt the model selection of an effect model as proposed in Kent et al. (2020) by using the strategy suggested in Wasserman and Roeder (2009). The following steps explain the ensuing effect model selection strategy.

Step 0. Initialization. Fix some $\alpha \in [0, 1]$. Let $\mathcal{I} \subset [p] := \{1, \ldots, p\}$ be the set of variables that are to be interacted with the treatment assignment variable.[1] Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be approximately equally sized, mutually disjoint subsets of the sample space, that is, $\bigcup_{i=1}^2 \mathcal{D}_i = [n]$, while $\bigcap_{i=1}^2 \mathcal{D}_i = \varnothing$. Let Λ_n be a finite set of pre-specified choices of the regularization tuning parameter λ. The parameter choices in Λ_n constitute the grid along which we perform cross-validation.[2]

Step 1. Fitting of a single effect model. Fix some $\lambda \in \Lambda_n$. For all $i = 1, \ldots, n$, we consider the linear logistic model identified by

$$\ln\left(\frac{\mathbb{P}[Y_i = 1 \mid W_i = w_i, X_i = x_i]}{1 - \mathbb{P}[Y_i = 1 \mid W_i = w_i, X_i = x_i]}\right) = \beta_0 + x_i^\top \beta + \gamma_0 w_i + \sum_{\{j : j \in \mathcal{I}\}} \gamma_j w_i x_{ij}, \tag{6}$$

where x_{ij} is the j-th element in vector x_i, whereas $\beta_0, \gamma_0 \in \mathbb{R}, \beta \in \mathbb{R}^p$, and $\gamma := \{\gamma_j\}_{j \in \mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ are coefficients. Collect the non-intercept coefficients in a vector of dimension $q = 1 + p + |\mathcal{I}|$, denoted $\vartheta := (\vartheta_1, \ldots, \vartheta_q) := (\beta, \gamma_0, \gamma)$. The q-dimensional

[1] The choice $\mathcal{I} = \varnothing$ is also permitted (i.e. no variable is interacted), but we decidedly do not reflect this possibility in our notation for the sake of notational brevity.

[2] See Friedman et al. (2010) for details on the construction of the grid Λ_n.

vector of independent variables that is associated with the coefficient vector ϑ is given by $z_i := (z_{i1}, \ldots, z_{iq}) := \big(x_i, w_i, \{w_i x_{ij}\}_{j\in\mathcal{I}}\big)$.

Define by $\mathcal{L}_{\mathcal{D}}(\beta_0, \vartheta)$ the log-likelihood function[3] corresponding to model (6), evaluated at some sample subset $\mathcal{D} \subset [n]$. We fit the model by solving

$$(\widetilde{\beta}_{0,\lambda}, \widetilde{\vartheta}_\lambda) = \arg \min_{(\beta_0,\vartheta)\in\mathbb{R}^{q+1}} \left\{ -\frac{1}{|\mathcal{D}_1|}\mathcal{L}_{\mathcal{D}_1}(\beta_0, \vartheta) + \lambda\Big((1-\alpha)\|\vartheta\|_2^2 + \alpha\|\vartheta\|_1\Big) \right\}, \quad (7)$$

that is, we obtain coefficient estimates $(\widetilde{\beta}_{0,\lambda}, \widetilde{\vartheta}_\lambda)$ by only using information from the observations in set $\mathcal{D}_1$. The estimates $(\widetilde{\beta}_{0,\lambda}, \widetilde{\vartheta}_\lambda)$ depend on the choice of tuning parameter λ; we make this dependence explicit in our notation by including a λ-subscript. Next, collect the retained variables in a suite

$$\widehat{S}_n(\lambda) := \Big\{ j \in [q] : \widetilde{\vartheta}_{j,\lambda} \neq 0 \Big\}.$$

Denote by

$$\widehat{L}(\lambda) := -\mathcal{L}_{\mathcal{D}_1}(\widetilde{\beta}_{0,\lambda}, \widetilde{\vartheta}_\lambda)$$

the empirical loss associated with tuning parameter λ. The empirical loss here corresponds to value of the negative log-likelihood evaluated at the observations in $\mathcal{D}_1$ and the the coefficient estimates $(\widetilde{\beta}_{0,\lambda}, \widetilde{\vartheta}_\lambda)$.

Step 2. Cross-validation. The optimal choice for the tuning parameter λ in (7) corresponds to the choice that minimizes the empirical loss, that is,

$$\widehat{\lambda} = \arg \min_{\lambda\in\Lambda_n} \widehat{L}(\lambda).$$

The *final model*, denoted $\widehat{S}_n$, contains the retained variables of the model associated with the optimal tuning parameter λ:

$$\widehat{S}_n := \widehat{S}_n(\widehat{\lambda}) \subset [q].$$

Step 3. Final model. For $i \in [n]$, recall that $z_i = (z_{i1}, \ldots, z_{ij}) = \big(x_i, w_i, \{w_i x_{ij}\}_{j\in\mathcal{I}}\big)$ is the q-dimensional vector that holds the independent variables in the linear logistic model (6) and that $\vartheta = (\vartheta_1, \ldots, \vartheta_q)$ is the associated vector of coefficients. Potentially, some of the coefficients in ϑ have been shrunk to zero in the final model $\widehat{S}_n$. Using a slight abuse of set theoretic notation, denote by

$$z_{S,i} = \Big\{ z_{ij} \in z_i : j \in \widehat{S}_n \Big\} \quad \text{and} \quad \vartheta_S = \Big\{ \vartheta_j \in \vartheta : j \in \widehat{S}_n \Big\} \quad (8)$$

the vector of independent variables that are retained in the final model $\widehat{S}_n$ and its associated coefficient vector, respectively. Both vectors are of dimension $s = |\widehat{S}_n|$, which corresponds to the number of retained variables in the reduced model.

[3] This log-likelihood function is given by $\mathcal{L}_{\mathcal{D}}(\beta_0, \vartheta) = \sum_{\{i:i\in\mathcal{D}\}} \Big(Y_i\big(\beta_0 + z_i^\top \vartheta\big) - \ln\big(1 + \mathrm{e}^{\beta_0 + z_i^\top \vartheta}\big)\Big)$.

On $\widehat{S}_n$ (that is, with these reduced vectors), we can subsequently perform logistic regression by using only the observations in the second sample subset, $\mathcal{D}_2$, to fit the final model. Concretely, we minimize the negative log-likelihood

$$(\widehat{\beta}_0, \widehat{\vartheta}_S) = \arg \min_{(\beta_0, \vartheta_S) \in \mathbb{R}^{s+1}} \left\{ - \sum_{\{i: i \in \mathcal{D}_2\}} \left(Y_i (\beta_0 + z_{S,i}^\top \vartheta_S) - \ln \left(1 + \exp \left(\beta_0 + z_{S,i}^\top \vartheta_S \right) \right) \right) \right\}.$$

The estimates $(\widehat{\beta}_0, \widehat{\vartheta}_S)$ are the estimated coefficients of the fitted final model. Elementary logistic regression theory allows us to compute a covariance matrix estimate $\widehat{\mathbb{V}}(\widehat{\vartheta}_{\widehat{S}_n})$ with which we can perform z-tests on the significance of the coefficients of the retained variables in $\widehat{S}_n$. We emphasize that all inference needs to be based on data in $\mathcal{D}_2$ to avoid a bias (see Wasserman and Roeder (2009) for details).

Step 4. Risk prediction with the final model. Given some $w \in \{0, 1\}$, the *risk prediction function* of the (final) effect model is, for all $i = 1, \ldots, n$, given by

$$\mathsf{risk}_i(w) := \widehat{\mathbb{P}}[Y_i = 1 | W_i = w, X_i = x_i] = F_{logistic}\Big(\widehat{\beta}_0 + z_{D,i}^\top \widehat{\vartheta}_S \Big), \tag{9}$$

where $z_{D,i}$ depends on w: recall that the original vector is given by $z_i = (z_{i1}, \ldots, z_{ij}) = \big(x_i, w_i, \{w_i x_{ij}\}_{j \in \mathcal{I}}\big)$, which can be expressed as $z_i = \big(x_i, w, \{w x_{ij}\}_{j \in \mathcal{I}}\big)$ to make the dependence on w explicit.

Observe that it may happen that treatment assignment $W_i = w_i$ is not among the retained variables in the final model $\widehat{D}_n$. This is problematic because if the treatment assignment is not retained, then the risk prediction function $w \mapsto \mathsf{risk}_i(w)$ in (9) is not defined. Kent et al. (2020) do not address this possibility. To overcome this potential issue, we could artificially add the treatment assignment to the retained variables. That is, if $j_w \in [q]$ is the index of the treatment assignment variable, we would define the final model by

$$\widehat{S}_n = \{j_w\} \cup \widehat{S}_n(\widehat{\lambda}).$$

However, this solution is potentially problematic from a statistical point of view [Indeed, this is an issue. Amdreas recommended to use the hierarchical elastic net to overcome this problem. In addition, he pointed out that we should check out a series of papers by ETH people whose post selection inference strategies are less prone to p-value hacking than the method of Wasserman and Roeder (2009).]

3.3 Estimation of Benefits

Suppose we have fitted a predictive model, either a risk model (as in Section 3.1 or an effect model (as in Section 3.2). Consider the risk prediction function $w \mapsto \mathsf{risk}_i(w), i = 1, \ldots, n$, associated with the fitted model. For the observed treatment status of individual i, $W_i = w_i$, we define the *regular risk* as $\mathsf{risk}_i^{\mathsf{reg}} = \mathsf{risk}_i(w_i)$. Furthermore, define the *reversed* treatment assignment by

$$W_i^{rev} = \begin{cases} 1 & \text{if } W_i = 0, \\ 0 & \text{if } W_i = 1. \end{cases}$$

Thereupon, at $W_i^{rev} = w_i^{rev}$, we define the *reverse risk* by $\mathsf{risk}_i^{\mathsf{rev}} = \mathsf{risk}_i(w_i^{rev})$. We estimate the absolute benefit in (1) by the *predicted absolute benefit*, $\widehat{\theta}_i$,

$$\widehat{\theta}_i = \mathsf{risk}_i - \mathsf{risk}_i^{\mathsf{rev}}$$

and the relative benefit in (2) by the *predicted relative benefit*, $\widehat{\theta}_i^{rel}$,

$$\widehat{\theta}_i^{rel} = \frac{\mathsf{risk}_i}{\mathsf{risk}_i^{\mathsf{rev}}}.$$

Effectively, estimation of the predictive benefits consists of estimating the potential outcomes. However, Chernozhukov et al. (2020) show that such approaches only give noisy (and therefore inconsistent) estimates of the causal parameter of interest. We use the predictive benefits to measure the strength of treatment effect heterogeneity, which will be discussed in the next section.

3.4 Measuring Heterogeneity

Suppose we have obtained predictive absolute and relative benefits, $\widehat{\theta}_i$ and $\widehat{\theta}_i^{rel}$, respectively, for $i = 1, \ldots, n$. Suppose furthermore that the observations $\{1, \ldots, n\}$ can be grouped into m groups which we index by $1, 2, \ldots, m$. Denote the group membership of observation i by $G_i \in \{1, \ldots, m\} = \mathcal{G}$. The grouping is usually done based on characteristics in the covariates X_i. Denote by

$$\begin{aligned} \widehat{\theta}_g &= |\{i \in [n] : G_i = g\}|^{-1} \sum_{\{i \in [n] : G_i = g\}} \widehat{\theta}_i \quad \text{and} \\ \widehat{\theta}_g^{rel} &= |\{i \in [n] : G_i = g\}|^{-1} \sum_{\{i \in [n] : G_i = g\}} \widehat{\theta}_i^{rel} \end{aligned} \tag{10}$$

the within-group estimates, for groups $g \in \mathcal{G}$. Since the within-group estimates are just means, one can use elementary t-tests for testing for differences between group estimates, which is intended as inference for the strength of treatment effect heterogeneity along various groups.

3.5 Measuring Calibration

In order to assess whether a fitted predictive model provides an accurate fit to the data, calibration plots are used. For this purpose, let the grouping in the previous section be done based on quantiles of the estimated baseline risk of having the event (see Section 3.1.1). Consider these groups as fixed for the remainder of this section. Moreover, calculate the group-level predicted benefits as in (10) on all groups $g \in \mathcal{G}$. A *calibration plot* plots the group-level predicted benefits against the *group-level observed benefits*. If the the group-level predicted benefits and group-level observed benefits roughly lie on a 45 degrees line, we say that the predictive model is well calibrated. In the following, we show how group-level observed benefits can be obtained.

3.5.1 Absolute Observed Benefit

For each group $g \in \mathcal{G}$ calculate the group-level *absolute* observed benefit $\widehat{aob}_g$ by

$$\begin{aligned}\widehat{aob}_g =& |\{i : G_i = g, W_i = 1\}|^{-1} \sum_{\{i:G_i=g,W_i=1\}} Y_i - \\ & |\{i : G_i = g, W_i = 0\}|^{-1} \sum_{\{i:G_i=g,W_i=0\}} Y_i.\end{aligned} \tag{11}$$

Since the group-level absolute observed benefit is essentially the difference in the average group-level mortality, we can view it as an estimate of the group-level average treatment effect.

Due to this "difference in means" structure, we can use elementary statistical theory to construct confidence intervals for the absolute observed benefit $\widehat{aob}_g$. Concretely, we can construct a *Welch t-test* for performing tests on $\widehat{aob}_g$. We use a Welch t-test rather than a classical t-test for the sake of generality: Unlike a classical t-test, Welch's test does not assume equal population variances of the two summands of $\widehat{aob}_g$ in equation (11).

3.5.2 Relative Observed Benefit

For each group $g \in \mathcal{G}$ calculate the group-level *relative* observed benefit $\widehat{rob}_g$ by the relative risk (or risk ratio)

$$\widehat{rob}_g = \frac{|\{i : G_i = g, W_i = 1\}|^{-1} \sum_{\{i:G_i=g,W_i=1\}} Y_i}{|\{i : G_i = g, W_i = 0\}|^{-1} \sum_{\{i:G_i=g,W_i=0\}} Y_i}. \tag{12}$$

Observe that the relative observed benefit corresponds to the ratio of the two summands that constitute the absolute observed benefit in equation (11) and can be viewed as a ratio of two proportions. The relative observed benefit estimates the theoretical group-level risk ratio

$$rr_g = \frac{\mathbb{P}(Y_1 = 1 | W_1 = 1, G_1 = g)}{\mathbb{P}(Y_1 = 1 | W_1 = 0, G_1 = g)}. \tag{13}$$

Constructing a confidence interval for the group-level relative observed benefit $\widehat{rob}_g$ is somewhat more involved than for the absolute observed benefit, because the ratio $\widehat{rob}_g$ usually does not asymptotically follow a normal distribution. Using the procedure suggested here, we construct confidence intervals for $\widehat{rob}_g$ by using a two-step procedure.

We first need some ancillary quantities. Within a given group $g \in \mathcal{G}$ let the constants

$$\begin{aligned} N_g^{(Y=1,W=1)} &= |\{i : G_i = g, Y_i = 1, W_i = 1\}|, \\ N_g^{(Y=0,W=1)} &= |\{i : G_i = g, Y_i = 0, W_i = 1\}|, \\ N_g^{(Y=1,W=0)} &= |\{i : G_i = g, Y_i = 1, W_i = 0\}|, \text{ and} \\ N_g^{(Y=0,W=0)} &= |\{i : G_i = g, Y_i = 0, W_i = 0\}| \end{aligned} \tag{14}$$

count the number of treated deaths, treated survivors, untreated deaths, and untreated survivors, respectively.[4] Similarly, let

$$N_g^{(W=1)} = |\{i : G_i = g, W_i = 1\}| \quad \text{and} \quad N_g^{(W=0)} = |\{i : G_i = g, W_i = 0\}| \tag{15}$$

count the number of treated and untreated samples, respectively, in group g. For a given group $g \in \mathcal{G}$, denote by

$$Z_g = \frac{N_g^{(Y=0,W=1)}}{N_g^{(Y=1,W=1)}} \Big/ N_g^{(W=1)} + \frac{N_g^{(Y=0,W=0)}}{N_g^{(Y=1,W=0)}} \Big/ N_g^{(W=0)} \tag{16}$$

the sum of the relative survival for the treatment and control group (normalized by the number of samples in each of these groups).

Finally, for a given group $g \in \mathcal{G}$, a corresponding relative observed benefit $\widehat{rob}_g$, a significance level $\alpha \in (0, 0.5)$, and a $(1-\alpha)$-quantile of the standard normal distribution $z_{1-\alpha/2}$, denote the real-valued interval $\mathcal{C}_g^{\log(rr)}$ by

$$\mathcal{C}_g^{\log(rr)} = \left[\log\left(\widehat{rob}_g\right) - z_{1-\alpha/2}\sqrt{Z_g},\ \log\left(\widehat{rob}_g\right) + z_{1-\alpha/2}\sqrt{Z_g}\right]. \tag{17}$$

The interval $\mathcal{C}'_g$ is a $(1-\alpha)$-confidence interval for $\log(rr_g)$. A simple exponential transformation of (17) yields the desired $(1-\alpha)$-confidence interval for rr_g:

$$\begin{aligned} \mathcal{C}_g^{rr} &= \left[\exp\left(\log\left(\widehat{rob}_g\right) - z_{1-\alpha/2}\sqrt{Z_g}\right), \exp\left(\log\left(\widehat{rob}_g\right) + z_{1-\alpha/2}\sqrt{Z_g}\right)\right] \\ &= \left[\widehat{rob}_g \exp\left(-z_{1-\alpha/2}\sqrt{Z_g}\right),\ \widehat{rob}_g \exp\left(z_{1-\alpha/2}\sqrt{Z_g}\right)\right]. \end{aligned} \tag{18}$$

3.6 Empirical Odds Ratio

The empirical odds ratio is often used as an additional performance measure and is discussed here for the sake of completeness. Define for the relevant group-level theoretical probabilities $p_g^{(1)} = \mathbb{P}(Y_1 = 1 | W_1 = 1, G_1 = g)$ and $p_g^{(0)} = \mathbb{P}(Y_1 = 1 | W_1 = 0, G_1 = g)$. Then, the relative risk in (13) can be written as $rr_g = p_g^{(1)}/p_g^{(0)}$. The theoretical group-level *odds ratio* is defined as

$$or_g = \frac{p_g^{(1)}/(1-p_g^{(1)})}{p_g^{(0)}/(1-p_g^{(0)})}. \tag{19}$$

Define now the empirical probabilities by

$$\hat{p}_g^{(1)} = \frac{N_g^{(Y=1,W=1)}}{N_g^{(Y=1,W=1)} + N_g^{(Y=0,W=1)}} \quad \text{and}$$
$$\hat{p}_g^{(0)} = \frac{N_g^{(Y=1,W=0)}}{N_g^{(Y=1,W=0)} + N_g^{(Y=0,W=0)}}.$$

[4] Usually these counters are represented in a 2×2 contingency table as in here.

Analogously to the theoretical group-level odds ratio or_g, the *empirical* odds ratio $\widehat{or}_g$ is given by

$$\widehat{or}_g = \frac{\hat{p}_g^{(1)}/(1-\hat{p}_g^{(1)})}{\hat{p}_g^{(0)}/(1-\hat{p}_g^{(0)})}. \tag{20}$$

Similarly to the relative observed benefit, we cannot immediately construct a confidence interval for the theoretical group-level odds ratio or_g, but need a two-step procedure. The variance of the estimator $\widehat{or}_g$ is given by

$$V_g = \frac{1}{N_g^{(Y=1,W=1)}} + \frac{1}{N_g^{(Y=0,W=1)}} + \frac{1}{N_g^{(Y=1,W=0)}} + \frac{1}{N_g^{(Y=0,W=0)}}. \tag{21}$$

For a given significance level $\alpha \in (0, 0.5)$ and a $(1-\alpha)$-quantile of the standard normal distribution $z_{1-\alpha/2}$, denote the real-valued interval $\mathcal{C}_g^{\log(or)}$ by

$$\mathcal{C}_g^{\log(or)} = \Big[\log\left(\widehat{or}_g\right) - z_{1-\alpha/2}\sqrt{V_g},\ \log\left(\widehat{or}_g\right) + z_{1-\alpha/2}\sqrt{V_g}\Big]. \tag{22}$$

The interval $\mathcal{C}_g^{\log(or)}$ is a $(1-\alpha)$-confidence interval for $\log(or_g)$. A simple exponential transformation yields he desired $(1-\alpha)$-confidence interval for or_g:

$$\begin{aligned} \mathcal{C}_g^{or} &= \left[\exp\Big(\log\left(\widehat{or}_g\right) - z_{1-\alpha/2}\sqrt{V_g}\Big), \exp\Big(\log\left(\widehat{or}_g\right) + z_{1-\alpha/2}\sqrt{V_g}\Big)\right] \\ &= \left[\widehat{or}_g \exp\Big(-z_{1-\alpha/2}\sqrt{V_g}\Big),\ \widehat{or}_g \exp\Big(z_{1-\alpha/2}\sqrt{V_g}\Big)\right]. \end{aligned} \tag{23}$$

References

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2020). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *arXiv preprint: arXiv1712.04802*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., et al. (2020). The predictive approaches to treatment effect heterogeneity (PATH) statement. *Annals of Internal Medicine*, 172(1):35–45.

Lehmann, E. L. and Romano, J. P. (1986). *Testing Statistical Hypotheses*. Wadsworth & Brooks/Cole, Pacific Grove, California, 2nd edition.

Rekkas, A., Paulus, J. K., Raman, G., Wong, J. B., Steyerberg, E. W., Rijnbeek, P. R., Kent, D. M., and van Klaveren, D. (2019). Predictive approaches to heterogeneous treatment effects: a systematic review. *medRxiv preprint: medRxiv:19010827*.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Annals of Statistics*, 37(5A):2178–2201.

Appendix

A (Regularized) Logistic Regression

Assume we are interested in predicting binary outcomes Y_i via q-dimensional random vectors Z_i, where $\{(Z_i, Y_i)\}_{i=1}^n$ is a random sample. Hence, since $Y_i|Z_i \sim \mathsf{Bernoulli}(p_{i,Z})$, we are effectively interested in estimating the parameter $p_{i,Z} = \mathbb{P}[Y_i = 1 \mid Z_i]$, where the subscript "$Z$" reminds us that $p_{i,Z}$ is a conditional probability.

We assume a logistic linear model, which assumes the linear identity

$$\ln\left(\frac{p_{i,Z}}{1-p_{i,Z}}\right) = \beta_0 + Z_i^\top \beta, \tag{24}$$

for all $i = 1, \ldots, n$, where $(\beta_0, \beta) = (\beta_0, \beta_1, \ldots, \beta_q) \in \mathbb{R}^{q+1}$ is some fixed vector of coefficients. Rearranging identifies the probability of interest, $p_{i,Z}$, as

$$p_{i,Z} = \left(1 + \exp\left(-\beta_0 - Z_i^\top \beta\right)\right)^{-1}.$$

We can fit the model (24) by *logistic regression*, which performs maximum likelihood estimation over the coefficients. The corresponding optimization problem can be shown to solve

$$(\widehat{\beta}_0, \widehat{\beta}) \in \arg\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left\{ -\frac{1}{n}\sum_{i=1}^n \left(Y_i(\beta_0 + Z_i^\top\beta) - \ln\left(1 + \mathrm{e}^{\beta_0 + Z_i^\top\beta}\right)\right)\right\}. \tag{25}$$

However, when there are more explanatory variables than observations, $q \geq n$, then the solution $(\widehat{\beta}_0, \widehat{\beta})$ may be degenerate. In this case, adding a regularization penalty to the objective function in (25) enforces sparsity in the solution and thereby renders the solution stable again. Concretely, given fixed $\lambda_n \geq 0$ and $\alpha \in [0, 1]$, *regularized logistic regression* solves

$$\begin{aligned}(\widehat{\beta}_0, \widehat{\beta}) \in \arg\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \Bigg\{ &-\frac{1}{n}\sum_{i=1}^n \left(Y_i(\beta_0 + Z_i^\top\beta) - \ln\left(1 + \mathrm{e}^{\beta_0 + Z_i^\top\beta}\right)\right) + \\ &+\lambda_n\Big((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\Big)\Bigg\}.\end{aligned} \tag{26}$$

Observe that for $\alpha = 0.5$, we obtain the elastic net penalty and for $\alpha = 1$, we obtain the Lasso penalty. We can find an appropriate choice of λ_n by cross-validation over a grid of candidate values for λ_n. See Friedman et al. (2010) for numerical details.

Suppose we have obtained an estimate $(\widehat{\beta}_0, \widehat{\beta})$ of (β_0, β) in the identity (24), either by regularized or non-regularized logistic regression. Then we can estimate the probabilities of interest $\mathbb{P}[Y_i = 1 \mid X_i]$ by

$$\widehat{p}_{i,Z} = \left(1 + \exp\left(-\widehat{\beta}_0 - Z_i^\top \widehat{\beta}\right)\right)^{-1}.$$