Probabilistic Index Models

Jan De Neve[1], Olivier Thas[1], Lieven Clement[2] and Jean-Pierre Ottoy[1]

[1]*Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium*

[2] *L-BioStat, K.U. Leuven, Belgium*

Abstract

We present a semiparametric statistical model for the probabilistic index which can be defined as $\mathrm{P}\{Y \leq Y'\}$, where Y and Y' are independent random response variables associated with covariate patterns $\boldsymbol{X}$ and $\boldsymbol{X}'$, respectively. A link function defines the relation between the probabilistic index and a linear predictor. Asymptotic normality of the estimators and consistency of the covariance matrix estimator is established through semiparametric theory. The model is illustrated on an example.

Keywords: Semiparametric Inference, Rank tests, Probabilistic Index, AUC regression, Effect Size

AMS subject classifications: 62J07

1 Introduction

The probabilistic index (PI) is defined in the present setting as the probability

$$\mathrm{P}\{Y < Y'\} + \frac{1}{2}\mathrm{P}\{Y = Y'\},$$

which is denoted by $\mathrm{P}\{Y \preccurlyeq Y'\}$ and where Y and Y' denote independent random variables. The PI acquired recently an increase of interest in Biostatistics. The PI is accepted as an informative and intuitive index for quantifying treatment effects, particularly when the treatment does not only act on the mean response. In a clinical trial for example, the effect of increased dose of a drug on a response can affect the mean, dispersion and even the tails of the density. The PI summarizes the effect on the shape of the distribution into a simple effect measure.

If the design of the experiment consists of only two groups, two treatments for example, the Wilcoxon-Mann-Whitney (WMW) test can be employed to analyse this type of data and provides an interpretation in terms of the PI. In the presence of confounders, extensions of WMW have been presented, see for example [1, 2, 3]. However, to our knowledge no method exists for quantifying the effect of, for example, a *continuous* predictor on the response in terms of the PI. In this paper we place the PI in a general regression context allowing to assess effects of predictors on the PI for a variety of designs. More specifically we propose the probabilistic index model (PIM)

$$\mathrm{P}\{Y \preccurlyeq Y' | \boldsymbol{X}, \boldsymbol{X}'\} = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \tag{1}$$

where $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ are independently and identically distributed random vectors where the response Y (Y') is associated with predictors $\boldsymbol{X}$ ($\boldsymbol{X}'$). Function m defines the relation between the probabilistic index and the predictors. To keep the method as general as possible, while still retaining an intuitive interpretation, the model estimation and inference are based on semiparametric

theory. Therefore no distributional assumptions are required, but the model imposes restrictions on the distribution.

Section 2 summarizes the asymptotic theory, Section 3 summarizes the relations with other models, while Section 4 illustrates the interpretation of the model on an example. Note that all these sections only briefly report on our results. Details can be found elsewhere, see [6]. Section 5 contains some conclusions.

2 Parameter Estimation and Statistical Inference

Define $\mathrm{I}\{Y \preccurlyeq Y'\} = \mathrm{I}\{Y < Y'\} + \frac{1}{2}\mathrm{I}\{Y = Y'\}$ in which $\mathrm{I}\{Y < Y'\}$ and $\mathrm{I}\{Y = Y'\}$ denote the usual indicator functions evaluated in the events $Y < Y'$ and $Y = Y'$, respectively. The PIM (1) can then also be written as

$$\mathrm{E}\left\{\mathrm{I}\{Y \preccurlyeq Y'\}|\boldsymbol{X}, \boldsymbol{X}'\right\} = \mathrm{P}\left\{Y \preccurlyeq Y'|\boldsymbol{X}, \boldsymbol{X}'\right\} = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = h(\boldsymbol{Z}^t\boldsymbol{\beta}), \tag{2}$$

where $h(\boldsymbol{Z}^t\boldsymbol{\beta})$ is used to stress that $m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta})$ depends on $\boldsymbol{X}$ and $\boldsymbol{X}'$ through the linear predictor $\boldsymbol{Z}^t\boldsymbol{\beta}$ with $\boldsymbol{Z}$ a vector which depends on elements of $\boldsymbol{X}$ and $\boldsymbol{X}'$ and with h the inverse of a link function, say g.

When $(Y_1, \boldsymbol{X}_1)$, $(Y_2, \boldsymbol{X}_2)$, …, $(Y_n, \boldsymbol{X}_n)$ denotes a sample of n i.i.d. random variables with joint density function f_{YX}, model formulation (2) suggests that the $\boldsymbol{\beta}$ parameter vector can be estimated using the set of *pseudo-observations* $I_{ij} = \mathrm{I}\{Y_i \preccurlyeq Y_j\}$ for all $i, j = 1, \ldots, n$. In particular, model (2) resembles a conditional moment semiparametric model in which the conditional mean of the pseudo-observations is specified. We therefore propose to estimate the parameters by solving the estimating equations

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{i,j} \boldsymbol{U}_{ij}(\boldsymbol{\beta}) = \sum_{i,j} \frac{\partial h(\boldsymbol{Z}_{ij}^t\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^t} \boldsymbol{V}^{-1}(h(\boldsymbol{Z}_{ij}^t\boldsymbol{\beta}))(I_{ij} - h(\boldsymbol{Z}_{ij}^t\boldsymbol{\beta})) = \boldsymbol{0}, \tag{3}$$

where $\boldsymbol{V}(h(\boldsymbol{Z}_{ij}^t\boldsymbol{\beta})) = \frac{1}{\nu}\mathrm{Var}\{I_{ij}|\boldsymbol{Z}_{ij}\}$, with ν a scale parameter. Let $\hat{\boldsymbol{\beta}}_n$ denote the solution of (3). This choice corresponds to the quasi-likelihood estimating equations as used, for example, in the analysis of binary longitudinal data, where they are also referred to as *generalized estimation equations* (GEE). In the present setting, however, the conditional mean does not refer to the mean of the conditional distribution of the response, but it refers to the mean of the pseudo-observations. Despite the close relation between our method of estimation and GEE, the asymptotic distributional properties of the estimator $\hat{\boldsymbol{\beta}}$ do not follow immediately from these theories, for the pseudo-observations I_{ij} possess a more complicated dependence structure than e.g. block independence as in clustered or longitudinal data. However the pseudo-observations possess the *sparse correlation* structure of [4] and therefore the semiparametric theory of [4] directly applies to our setting. Theorem 1 summarizes the most important distribution theory results for the PIM.

Theorem 1 (Asymptotic normality) *Consider the PIM (2) with predictors $\boldsymbol{Z}_{ij}$ taking values in a bounded subset of $\mathbb{R}^p$. Assume*

A1 the pseudo-observations are sparsely correlated;

A2 the link function g and the variance function $\boldsymbol{V}$ have three continuous derivatives;

A3 the true parameter $\boldsymbol{\beta}_0$, is in the interior of a convex parameter space;

A4 there exists a vector $\boldsymbol{W}$ and positive definite matrix $\boldsymbol{T}$ such that

$$n^{-2}\sum_{i,j}\boldsymbol{Z}_{ij} \to \boldsymbol{W} \text{ and } n^{-2}\sum_{i,j}\boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^t \to \boldsymbol{T};$$

A5 $\limsup n^{-1}\mathrm{Var}\Big(\sum_{i,j} I_{ij}\Big) > 0.$

Then, as $n \to \infty$, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ *converges in distribution to a multivariate Gaussian distribution with zero mean and some positive definite variance-covariance matrix* $\boldsymbol{\Sigma}$.

Theorem 2 (Consistent variance estimator) *Under the regularity conditions of Theorem 1, the variance-covariance matrix* $\boldsymbol{\Sigma}$ *can be consistently estimated by the sandwich estimator*

$$n\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_n} = n\Big(\sum_{i,j} \frac{\partial \boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}^t}\Big)^{-1}\Big(\sum_{i,j}\sum_{k,l} \phi_{ijkl}\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)\boldsymbol{U}_{kl}^t(\hat{\boldsymbol{\beta}}_n)\Big)\Big(\sum_{i,j} \frac{\partial \boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}_n)}{\partial \boldsymbol{\beta}^t}\Big)^{-1},$$

where the indicator ϕ_{ijkl} *is defined as* $\phi_{ijkl} = 1$ *if* I_{ij} *and* I_{kl} *are correlated and* $\phi_{ijkl} = 0$ *otherwise.*

Note that in (3) we use the independence working correlation matrix, therefore our proposed estimator will not be the most efficient one.

3 Relation with other Models

In this section we start from well known models and construct the induced PIM. This gives insights in how the predictor $\boldsymbol{Z}$ should be constructed from $\boldsymbol{X}$ and $\boldsymbol{X}'$ for these cases.

It can be shown that the following relations hold.

- If $Y = \boldsymbol{X}^t\boldsymbol{\alpha} + \varepsilon$ with $\varepsilon \sim \mathrm{N}(0,\sigma^2)$ then

$$\mathrm{P}\left\{Y \preccurlyeq Y'|\boldsymbol{X},\boldsymbol{X}'\right\} = \Phi\left((\boldsymbol{X}'-\boldsymbol{X})^t\boldsymbol{\beta}\right) \text{ with } \boldsymbol{\beta} = \frac{\boldsymbol{\alpha}}{\sqrt{2}\sigma},$$

  where Φ denotes the standard normal distribution function.

- If the proportional hazards model hold: $\lambda(y|\boldsymbol{X}) = \lambda_0(y)\exp(\boldsymbol{X}^t\boldsymbol{\beta})$ with $\lambda_0(y)$ the baseline hazards function then

$$\mathrm{P}\left\{Y \preccurlyeq Y'|\boldsymbol{X},\boldsymbol{X}'\right\} = \mathrm{expit}\left((\boldsymbol{X}-\boldsymbol{X}')^t\boldsymbol{\beta}\right),$$

  with $\mathrm{expit}(x) = \exp(x)/(1+\exp(x))$.

- Consider the 2-sample case where the predictor X takes on 2 values, say $X = 0$ or $X = 1$ and consider the model

$$\mathrm{P}\{Y \preccurlyeq Y'|X,X'\} = g^{-1}\{(X'-X)\beta\}.$$

  Then the estimator of β, say $\hat{\beta}$, is proportional to the Wilcoxon-Mann-Whitney statistic.

These results indicate that $\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$ is a natural choice for the covariate function.

4 Example

The Mental Health Study (MHS) is a study of mental health for a random sample of 40 adult residents of Alachua County, Florida. See [5], p.185 for more information. The response variable is Mental Impairment (MI), which is ordinal with categories 1 (well), 2 (mild symptom formation), 3 (moderate

symptom formation) and 4 (impaired). Besides the mental impairment, the life index (LI) and the socioeconomic status (SES) are also reported. The SES is a binary variable coded as 0 (low SES) and 1 (high SES). The LI is a composite measure that quantifies the severity and the number of important life events such as birth of a child, death in family, divorce, etc. One of the objectives of the study is to assess whether the SES has an effect on MI. As the average MI score has no clear interpretation, [5] analysed the data with a cumulative logistic regression model. Here we analyse the data in terms of the probabilistic index. As it is believed that the LI may be a potential confounder, we propose to analyse the MHS data with the PIM

$$\text{logit}\left(\text{P}\left\{MI \preccurlyeq MI'\right\}\right) = \beta_1(SES' - SES) + \beta_2(LI' - LI). \tag{4}$$

A basic goodness-of-fit plot revealed no lack-of-fit (plot not shown). The model estimates are given by $\hat{\beta}_1 = -0.74$ (SE: 0.34, p-value: 0.03) and $\hat{\beta}_2 = 0.20$ (0.07, 0.006). The PIM analysis shows that, at the 5% level of significance, SES and LI have significant effects on the MI score in terms of the PI. With $\hat{\beta}_1 = -0.74$ we conclude that of people with equal LI, someone with a high SES has an estimated probability of $\text{expit}(-0.74) = 32\%$ to have a larger MI score than someone with a low SES and a 95% confidence interval is given by $[20\%, 48\%]$. People with a low SES are thus more likely to be mentally impaired than others with a high SES, while all having the same LI. The effect of the LI on MI can be estimated by the probability $\text{expit}(\hat{\beta}_2)$. In particular, among persons with the same SES, those with a LI of one unit smaller than the LI of another group of people, have a smaller MI score with estimated probability $\text{expit}(0.2) = 55\%$ with a 95% confidence interval of $[51\%, 59\%]$. Thus, the larger the LI, the more likely that someone is mentally impaired.

5 Conclusion

In this short communication some of the results related to probabilistic index models are presented, see [6] for more information. The main focus lies on the use of the probabilistic index, which has an intuitive interpretation, as a statistic to quantify treatment effects and this within a semiparametric regression framework.

References

[1] Brumback, C.L. Pepe, M.S. and Alonzo, T.A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in medicine* 25, 575–590.

[2] Dodd, L.E. and Pepe, M.S. (2003). Semiparametric Regression for the Area Under the Receiver Operating Characteristic Curve. *Journal of the American Statistical Association* 98(462), 409–417.

[3] Tian, L. (2008). Confidence intervals for $P(Y_1 > Y_2)$ with normal outcomes in linear models. *Statistics in medicine*, 27 4221–4237.

[4] Lumley, T. and Hamblett, N.M. (2003). Asymptotics for marginal generalized linear models with sparse correlations. *Technical Report 207, UW Biostatistics Working Paper Series, University of Washington.*

[5] Agresti, A. (2007) An introduction to categorical data analysis. *Hoboken, NJ, USA: Wiley.*

[6] Thas, O. De Neve, J. Clement, L. and Ottoy, J.P. Probabilistic Index Models. *Submitted to JRSS-B.*