A tutorial on probabilistic index models: regression models for the effect size P (Y1 < Y2)

Maarten De Schryver[1] & Jan De Neve[2]

[1] Department of Experimental Clinical and Health Psychology, Ghent University, Belgium

[2] Department of Data Analysis, Ghent University, Belgium

Author note

This research was supported by grant BOF09/01M00209 of Ghent University to Jan De Houwer. Correspondence concerning this paper should be addressed to Jan De Neve, Department of Data Analysis, Ghent University, Henri Dunantlaan 2, B-9000 Ghent (Belgium), Email: Jan.DeNeve@UGent.be, Phone: +32-(0)9 264 64 87.

Abstract

The probabilistic index (PI), also known as the probability of superiority or the common language effect size, refers to the probability that the outcome of a randomly selected subject exceeds the outcome of another randomly selected subject, conditional on the covariate values of both subjects. This summary measure has a long history, especially for the two-sample design where the covariate value typically refers to one of two treatments. Despite some of the attractive features of the PI, it is often not used beyond the two-sample design. One reason is the lack of a flexible regression framework that embeds the PI and that allows the user to estimate it for more complicated designs. However, Thas, De Neve, Clement, and Ottoy (2012) recently developed such a regression framework, named Probabilistic Index Models (PIMs). In this tutorial we provide an introduction to PIMs where we discuss several theoretical properties, motivate why we think PIMs could be useful for behavioral sciences, and illustrate how it can be used in practice using the R package **pim**.

Keywords: common language effect size, nonparametric, probability of superiority, regression, rank tests, semiparametric probability of superiority

A tutorial on probabilistic index models: Regression models for the effect size P (Y1 < Y2)

Probabilistic index models (PIMs) are a class of semiparametric regression models in which the probabilistic index (PI) is modeled as a function of covariates. The PI refers to the probability that the outcome of a randomly selected subject exceeds the outcome of another randomly selected subject, conditional on the covariate values of both subjects. Let Y denote the univariate outcome and $\boldsymbol{X}$ the p-dimensional vector of covariates. If $(Y_i, \boldsymbol{X}_i^T)$ denotes the observation of subject i and $(Y_j, \boldsymbol{X}_j^T)$ that of subject j, then the PI is given by $\mathrm{P}\left(Y_i < Y_j \,\middle|\, \boldsymbol{X}_i, \boldsymbol{X}_j\right)$.

Throughout this article we consider an example case study to make the notation more concrete. Let Y denote the Beck Depression Inventory (BDI) II depression score (range 0 – 63, with lower scores indicating less severe depression) after treatment, where patients were randomized to an innovative therapy (dummy coded as $X = 0$) or a conventional therapy ($X = 1$). We deliberately choose this dummy coding (with the innovative therapy as the reference) because the PI $\mathrm{P}\left(Y_i < Y_j \,\middle|\, X_i = 0, X_j = 1\right)$ then gives the probability that a randomly selected patient of the innovative therapy group (subject i) will have a better (thus lower) BDI score than a randomly selected patient of the conventional therapy group (subject j). A PI exceeding 50% states that subjects more often have lower BDI scores when receiving innovative treatment as compared to subjects receiving conventional treatment. In other words, it is more likely that a patient from the innovative treatment will be better off than a patient from the conventional treatment. For a PI less than 50%, the opposite holds: It is less likely that a patient from the innovative treatment will be better off than a patient from the conventional treatment. Or equivalently, it is more likely that a patient from the conventional treatment will be better off than a patient from the innovative treatment. This follows from $\mathrm{P}\left(Y_i > Y_j \,\middle|\, X_i = 0, X_j = 1\right) = 1 - \mathrm{P}\left(Y_i < Y_j \,\middle|\, X_i = 0, X_j = 1\right)$, so if $\mathrm{P}\left(Y_i < Y_j \,\middle|\, X_i = 0, X_j = 1\right) < 0.5$ then $\mathrm{P}\left(Y_i > Y_j \,\middle|\, X_i = 0, X_j = 1\right) > 0.5$. A PI of 50% indicates that a patient of the innovative treatment is as likely to be better or worse as compared to a patient from the conventional treatment: $\mathrm{P}\left(Y_i > Y_j \,\middle|\, X_i = 0, X_j = 1\right) = \mathrm{P}\left(Y_i < Y_j \,\middle|\, X_i = 0, X_j = 1\right) = 0.5$.

The PI is the effect measure associated with the Wilcoxon–Mann–Whitney test (also known as the Mann–Whitney test or the Wilcoxon-rank-sum test; Wilcoxon, 1945; Mann and Whitney, 1947) and has been studied extensively since the introduction of the Wilcoxon–Mann–Whitney test. We provide a brief and selective review of the PI where we make a distinction between advantages and disadvantages of using the PI as an effect measure.

Cliff (1993) argued that much behavioral data can only be given an ordinal-scale status. That is, the interval-scale status of some given data requires nontrivial empirical support for the assertion that nominally equal intervals at different points on the scale are equal. The majority of psychometrically defined variables do not meet this criterion. The PI is an appropriate effect measure because it provides an ordinal answer to an ordinal question. As such, the PI is relevant for ordinal data, because it only exploits the order of the variables (i.e. $Y_i < Y_j$) and not the magnitude of their differences (i.e. $Y_i - Y_j$). Acion, Peterson, Temple, and Arndt (2006) argued that the PI can be a better effect size measure when distributions are highly skewed than one using a standardized difference, because shifts expressed in standard deviations might then be difficult to understand. This is discussed in more detail in the next section. Ruscio (2008) illustrated that, unlike Cohen's d or the point-biserial correlation, the PI estimator is robust to base rates (i.e. unequal sample sizes). Because the PI is unaffected by monotone transformations, it can be considered a relevant effect measure when the observed outcome is monotonically related to an underlying latent variable (Grissom & Kim, 2001). For reaction time experiments, for example, the latency can be considered as a surrogate for some unobserved variable such as processing difficulty (Cliff, 1993). Instead of modeling the reaction time, one could then also model the reciprocal of time (i.e. the response frequency). The PI will not be affected by this reciprocal transformation, because $\mathrm{P}\,(Y_i < Y_j) = \mathrm{P}\,(Y_i^{-1} > Y_j^{-1})$. Another attractive feature of the PI is its robustness against outliers. Because it only exploits the order of the outcomes, extreme large or small values have limited impact.

In a series of papers, Senn (1997, 2006, 2011) discussed several limitations of the PI as an effect measure in clinical trials. The PI is an ordinal measure and does not provide information on the magnitude of the difference between two populations. Thus, for interval-scale data, the PI does not fully exploit the information available in the data. Further, the PI can be easily misinterpreted. For instance, for the example case study, the PI does not give the probability that a single patient will benefit from the innovative treatment as compared to the conventional treatment. Instead, the PI compares two different subjects, each from one treatment. The PI is also unable to distinguish between distributions with minimal overlap. Several of the properties of the PI (advantages and disadvantages) are discussed in more detail in the next section.

For the two-sample problem, P $\left(Y_i < Y_j \,\middle|\, X_i = 0, X_j = 1\right)$ has been given many names: the measure of stochastic superiority, the probability of superiority, the common language effect size, the dominance statistic, the nonparametric treatment effect, the relative treatment effect, the individual exceedance probability, and the probabilistic index, among others. We choose the latter, a term coined by Acion et al. (2006), while acknowledging that this name is not optimal (this also holds true for the other names), because other probability based effect measures can be constructed and therefore, the probability P $\left(Y_i < Y_j \,\middle|\, X_i = 0, X_j = 1\right)$ can be considered as *a* PI and not necessarily *the* PI.

The vast majority of articles on the PI focus on the two-sample design because an estimator of the PI can then be obtained from the Wilcoxon–Mann–Whitney test. Some authors have extended the estimation to more complicated designs. Tian (2008), for example, developed a parametric regression model for the PI assuming a normal linear regression model. Brumback, Pepe, and Alonzo (2006) developed a semiparametric model by using methods for receiver operating characteristic curve regression analysis to accommodate the Wilcoxon–Mann–Whitney test for covariate adjustment (Dodd & Pepe, 2003). Their methodology is, however, still restricted to two-sample designs, and does not allow quantification of the effect of a continuous covariate on the outcome in terms of the PI. Thas et al. (2012) introduced a class of regression models, named Probabilistic Index Models (PIMs), where they model the PI directly as a function of the covariates and this in a semiparametric fashion. This methodology allows estimating the PI for a variety of designs, including designs with multiple and/or continuous covariates. This allows, for example, to estimate the PI for comparing two treatments (e.g. innovative versus conventional therapy) within subpopulations (e.g. patients of a certain age). In De Neve and Thas (2015) it was further shown that many well-known nonparametric rank tests (e.g. Wilcoxon–Mann–Whitney, Kruskal–Wallis, Friedman rank tests) can be embedded in the PIM-framework in a similar fashion as how *t*- and *F*-tests can be embedded in a linear regression model.

Since the introduction of PIMs in Thas et al. (2012), several follow-up articles have been written (De Neve, Thas, & Ottoy, 2013a; De Neve, Thas, Ottoy, & Clement, 2013b; De Neve, Meys, Ottoy, Clement, & Thas, 2014; De Neve & Thas, 2015; Vermeulen, Thas, Vansteelandt, 2015; Amorim, Thas, Vermeulen, Vansteelandt, & De Neve, 2017). These articles are not tutorials but instead focus on applications of PIMs in the health and biological sciences. In this article we provide an introduction to PIMs with specific focus on the behavioral sciences. We discuss both theoretical and practical results and provide R code in the supplementary material on how the R package **pim** (Meys, De Neve, Sabbe, & Amorim, 2017; R Core Team, 2017) can be used.

The remainder of the paper is organized as follows. We first discuss several aspects of the PI in the two-sample design. Some of these results have been discussed in the literature. Nevertheless, we repeat them here as they are relevant for understanding PIMs. We then introduce PIMs by considering a univariate continuous covariate, before extending it to the multi-variable setting. We illustrate the relationship between PIMs and several other models, including normal linear regression and the Wilcoxon-Mann-Whitney test. We discuss goodness-of-fit assessment and illustrate the method on a case study.

The Probabilistic Index for the Two-sample Design

We start by comparing the PI with the mean difference and the standardized mean difference when the outcome follows a normal or a skewed distribution.

The (Standardized) Mean Difference and the Probabilistic Index under Normality

For the normal distribution, the PI is an effect measure that captures effects on both the mean and the variance. We consider the context of the introduction example, with Y the BDI score and X the treatment, where each patient is assigned to a single treatment. By Y_{IT} (Y_{CT}) we denote the outcome of the innovative (conventional) therapy so that we can write the PI compactly as P $(Y_{CT} < Y_{IT})$. We further assume that both outcomes are independent and normally distributed:

$Y_{IT} \sim N(\mu_{IT}, \sigma_{IT}^2)$, and $Y_{CT} \sim N(\mu_{CT}, \sigma_{CT}^2)$. To summarize the association between Y and X, we can look at the mean difference $\mu_{CT} - \mu_{IT}$, or its standardized version:

$$\delta = \frac{\mu_{CT} - \mu_{IT}}{\sqrt{(\sigma_{CT}^2 + \sigma_{IT}^2)/2}} .$$

This definition of δ, as proposed in Cohen (1988), allows the variances of the groups to differ and simplifies to $(\mu_{CT} - \mu_{IT})/\sigma$, when the variances are equal.

The PI and the standardized difference now have the following relationship:

$$\text{P}\ (Y_{IT} < Y_{CT}) = \Phi(\delta/\sqrt{2})\ , \qquad (1)$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. We refer to the supplementary material for more information.

We consider three examples to illustrate the three effect measures: the mean difference, the standardized mean difference δ and the PI. None of these examples are intended to promote one measure over the other; instead, they are meant to demonstrate their differences and similarities.

Example 1: constant mean difference, varying δ, and PI. Figure 1 (panels a and b) provides artificial data for BDI depression scores for two groups of patients receiving the different treatments (innovative or conventional). Within each treatment cohort there are two groups of patients: patients that receive antidepressants (panel a) and patients that do not receive antidepressants (panel b). We first consider panel a. The mean difference is 5 and $\delta = 1.05$. It is clear that the treatment affects both the mean and the variance of the distribution. The standardized mean difference captures both effects in a single summary measure. The interpretation of δ might, however, be more difficult to understand as compared to the mean difference: It states that the mean BDI score in the innovative treatment group will be 1.05 standard deviations lower as compared to the mean BDI score in the conventional treatment group. Thinking in units of standard deviations might be difficult, because it refers to a statistical unit instead of a clinically useful unit (Acion et al., 2006). Cut-off values can however help to improve the understanding of the magnitude of δ. The rules proposed in Cohen (1988) are often (explicitly or implicitly) used, where an effect is small if $\delta = 0.20$, medium if $\delta = 0.50$, and large if $\delta = 0.80$. It is important to realize that this classification is derived assuming both normality and constant variances and might not be appropriate when these assumptions are violated. We demonstrate this for a skewed distribution at the end of this section.

The third effect measure, the PI, accounts for the change in variability, while retaining an understandable interpretation in terms of a probability. For panel a of Figure 1, it holds that $\mathrm{P}\,(Y_{IT} < Y_{CT}) = \Phi(\delta/\sqrt{2}) = 77\%$, that is, there is a 77% chance that a patient of the innovative treatment group will have a lower BDI score as compared to a patient of the conventional treatment group. The interpretation of the PI might be easier to communicate than δ, because it is a probability that a patient from the innovative group will be better off as compared to a patient from the conventional group and it is not expressed in units of standard deviation.

The importance of standardization becomes even more apparent when looking to panel b of Figure 1. It is visually clear that the effect of the therapy is different as compared to panel a. The mean difference, however, does not pick this up: It remains 5. In contrast, the standardized difference ($\delta = 0.53$) and the PI ($\mathrm{P}\,(Y_{IT} < Y_{CT}) = 65\%$) pick up this change and show a decreased effect due to an increase in variability.

Example 2: constant PI and varying mean difference and δ. The interpretation of the PI can be considered an attractive property. However, as mentioned in the introduction, it also has limitations. It is, for example, not able to distinguish between normal distributions that show minimal overlap. Panels c and d of Figure 1 illustrate this. The mean difference for the patients receiving antidepressant drugs is 10, while for patients not receiving antidepressants this is 20. For the standardized differences, effects are $\delta = 5$ and $\delta = 10$. Both the mean difference and the standardized difference indicate a larger treatment effect for the no antidepressant group. The PI, on the other hand, is not capable of quantifying this difference because it is approximately 1 for both groups because $\Phi(5/\sqrt{2}) \approx 99.98\%$ and $\Phi(10/\sqrt{2}) \approx 100\%$. Thus, from the moment that all the outcomes of one group exceed most of the outcomes of another group, the PI will be close to one, independent of how far both distributions are separated.

Example 3: constant δ and PI and varying mean difference. Panels e and f of Figure 1 show mean differences that are not equal: 4 for panel e and 8 for panel f. Both $\delta = 1.33$ and the PI

P $(Y_{IT} < Y_{CT}) = 83\%$ are constant however, because the change in variability between both panels cancels out the change in difference in means.

The three examples illustrate that the mean difference, the standardized mean difference and the PI are three different measures to quantify treatment effects, each with their own strengths and weaknesses.

The (Standardized) Mean Difference and the Probabilistic Index for Skewed Distributions

When we leave the strict assumption of normality and consider skewed distributions, the difference between the PI and δ becomes more apparent. To illustrate this, we consider the log-normal (LN) distribution $Y_{IT} \sim LN(\mu_{IT}, \sigma_{IT}^2)$ and $Y_{CT} \sim LN(\mu_{CT}, \sigma_{CT}^2)$. For the log-normal distribution, the mean difference equals

$$\mathrm{E}(\mathrm{Y}_{CT}) - \mathrm{E}(\mathrm{Y}_{IT}) = \exp\left(\mu_{CT} + \frac{\sigma_{CT}^2}{2}\right) - \exp\left(\mu_{IT} + \frac{\sigma_{IT}^2}{2}\right),$$

and the standardized mean difference equals

$$\delta = \sqrt{2}\,\frac{\exp\left(\mu_{CT}+\frac{\sigma_{CT}^2}{2}\right)-\exp\left(\mu_{IT}+\frac{\sigma_{IT}^2}{2}\right)}{\sqrt{\exp(\sigma_{CT}^2-1)\exp(2\mu_{CT}+\sigma_{CT}^2)+\exp(\sigma_{IT}^2-1)\exp(2\mu_{IT}+\sigma_{IT}^2)}}.$$

The log-normal distribution arises from exponentiating normal variables and because the PI is not affected by the exponentiation it follows that

$$\mathrm{P}\,(Y_{IT} < Y_{CT}) = \Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sqrt{\sigma_{CT}^2+\sigma_{IT}^2}}\right);$$

we refer to the supplementary material for more information.

This demonstrates that the PI is not just a transformation of δ to a probability. It is instead an effect measure on its own. Figure 2 illustrates this numerically. Panel a displays two normal distributions with variance 1 and mean difference 0.5, resulting in $\delta = 0.5$ and P $(Y_{IT} < Y_{CT}) = 63.8\%$. Panel b shows two log normal distributions with variance 4.7 and a mean difference of 1.09, so that $\delta = 0.5$ and P $(Y_{IT} < Y_{CT}) = 77.5\%$. Both standardized differences are equal, but from the panels it is clear that log normal distributions are more separated from each other. This visual difference is numerically translated in the PI, but not in the standardized difference. The figures further demonstrate that classifying $\delta = 0.5$ as a medium effect (Cohen, 1988) does not translate over distributions: Both panels would be classified as a medium effect, while panel b shows less overlap.

Interpretation

The PI discussed in this article gives a probabilistic statement about the relative ordering of the outcomes of two independent subjects from the two treatment groups (innovative or conventional), and this is not the same as comparing the same subject under the two treatments. To illustrate this, we resume the depression example, where Y_{IT} denotes the depression score under innovative treatment, and Y_{CT} the depression score under conventional treatment.

To simplify the exposition, we assume that $Y_{IT} \sim N(\mu_{IT}, \sigma_{IT}^2)$, and $Y_{CT} \sim N(\mu_{CT}, \sigma_{CT}^2)$, so that $Y_{IT} - Y_{CT} \sim N(\mu_{IT} - \mu_{CT}, \sigma^{*2})$, where $\sigma^{*2} = \mathrm{Var}(Y_{IT} - Y_{CT})$. It follows that:

$$\mathrm{P}\,(Y_{IT} < Y_{CT}) = P(Y_{IT} - Y_{CT} < 0) = \Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sigma^*}\right).$$

If Y_{CT} and Y_{IT} refer to outcomes of two independent subjects, then $\sigma^{*2} = \sigma_{CT}^2 + \sigma_{IT}^2$, so that the PI equals

$$\Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sqrt{\sigma_{CT}^2 + \sigma_{IT}^2}}\right).$$

However, if Y_{CT} and Y_{IT} refer to outcomes of the same subject, then $\sigma^{*2} = \sigma_{CT}^2 + \sigma_{IT}^2 - 2\mathrm{Cov}(Y_{CT}, Y_{IT})$, where the covariance is different from zero because the outcomes are not independent (they are measured on the same subject). The PI now becomes

$$\Phi\left(\frac{\mu_{CT} - \mu_{IT}}{\sqrt{\sigma_{CT}^2 + \sigma_{IT}^2 - 2\mathrm{Cov}(Y_{CT}, Y_{IT})}}\right).$$

Both PIs are not the same, and if each subject only receives one of the two treatments we can estimate the former, but it will in general not be possible to estimate the latter. This is fundamentally different when we look at the mean difference as an effect size. In both settings we have

$$E(Y_{IT}) - E(Y_{CT}) = \mu_{IT} - \mu_{CT}.$$

Consequently, when we choose the mean difference as effect measure and for this simple setting, we do not need to make a distinction whether we compare the same subject or if we compare two different subjects, the effect size is the same. We refer to Morris and DeShon (2002) for a more detailed discussion on effect sizes for repeated measures and independent designs and to Fay, Brittain, Shih, Follmann, and Gabriel (2018) for a detailed discussion on the causal interpretation of the PI in randomized experiments.

Gaining Power by Exploiting Order

A popular estimator of the PI is the one that forms the basis of the non-parametric Wilcoxon–Mann–Whitney (WMW) test. If we have n_{IT} subjects receiving innovative therapy and n_{CT} subjects receiving conventional therapy, the PI can be unbiasedly estimated via the Mann–Whitney statistic

$$MW = \frac{1}{n_{IT}n_{CT}}\sum_{i=1}^{n_{IT}}\sum_{j=1}^{n_{CT}} \mathrm{I}(Y_{IT,i} < Y_{CT,j}), \tag{2}$$

where $\mathrm{I}(\cdot)$ denotes the indicator function for which $\mathrm{I}\big(Y_{IT,i} < Y_{CT,j}\big) = 1$, if $Y_{IT,i} < Y_{CT,j}$, and $\mathrm{I}\big(Y_{IT,i} < Y_{CT,j}\big) = 0$, otherwise (ties are addressed later). Each subject of the innovative treatment is thus compared to each subject of the conventional treatment. A one is assigned if the outcome of the innovative treatment is the lower of the two, and a zero otherwise. The statistic MW then equals the mean of these binary values. Under the null hypothesis of equal distributions, it holds that $\mathrm{P}(Y_{IT} < Y_{CT}) = 0.5$ and, in the absence of ties, that $\mathrm{Var}(MW) = (n_{IT} + n_{CT} + 1)/12n_{IT}n_{CT}$ (Hollander, Wolfe and Chicken, 2013). Standardizing the Mann-Whitney statistic then results in the in the WMW test statistic:

$$\sqrt{\frac{12n_{IT}n_{CT}}{n_{IT}+ n_{CT}+1}}\,(MW - 0.5).$$

This expression makes it clear that the PI is the effect size associated with the WMW test: The test statistic is composed of an unbiased estimator of the PI. The null hypothesis of the WMW test states that both distributions are equal, and the (two-sided) alternative that $\mathrm{P}(Y_{IT} < Y_{CT}) \neq 0.5$ (Hollander et al, 2013).

At first sight, it might seem that the WMW test will lose power by only considering the relative ordering of the outcomes and by ignoring the magnitude of the differences. However, performing a hypothesis test based on Equation 2 can lead to a substantial gain in power as compared to a test based on the difference in sample means (Hollander et al., 2013). Table 1, taken from Van der Vaart (1998) and Hollander et al. (2013), shows the asymptotic relative efficiency (ARE) of the two-sample t-test relative to the WMW test when both groups only differ in their means; i.e. both distributions are identical, up to a location shift. The ARE is the limit of the fraction of two sample sizes. In the numerator we have the sample size of the t-test while the denominator gives the sample size the WMW test requires to have the same power as the t-test. When the distributions are normal (the data of both groups follow a normal distribution with the same variance, but different means), the ARE is less than one. This implies that the t-test is superior to the WMW test: For large samples, using the t-test requires only 95% observations of the WMW test to achieve the same level of power. Notice that this superiority is rather modest.

There are, however, several distributions for which the ARE exceeds one, implying that the WMW test is (substantially) superior. If the data in both groups are exponentially distributed, the t-test needs three times as many observations as the WMW test to achieve the same power. It is worth mentioning that the superior performance of the t-test is bounded: It cannot perform better than when data are samples from the (artificial) density $f(y) = \max(1 - y^2, 0)$ (Lehmann, 2004). In contrast, the superiority of the WMW test is unbounded when data are coming from the heavy-tailed Cauchy distribution.

To illustrate that the WMW test can also be superior in small samples (thus not relying on results displayed in Table 1 which are derived assuming large samples), we consider a Monte-Carlo simulation study to compare the power function of both tests. Figure 2 gives the power (approximated based on 10000 Monte-Carlo simulations) for balanced two-sample designs (with 20 or 40 observations per group), where the mean difference between both groups equals a half standard deviation and the variances are equal. A permutation null distribution is used for both tests so that deviations from normality do not invalidate the statistical inference. The ARE considers the relative sample sizes needed for both tests to achieve the same power and only holds for large samples. Figure 2, on the other hand, gives the power of the tests for fixed sample sizes and holds for small samples. We can see that even for small samples, the WMW test can outperform the t-test in terms of power. For the normal and the uniform distributions, the t-test is slightly more powerful, while for the other distributions the WMW test is more powerful. For the exponential distribution, the difference is most pronounced: The WMW test with 40 observations per group has a power of more than 85% to detect a difference of a half standard deviation, while for the t-test the power is approximately only 60%.

Extension to a Regression Context

Equation 2 can only be used to estimate the PI for a two-sample design where the covariate X is binary. If we want to estimate the PI when X denotes a continuous covariate or when there are multiple covariates, we need to embed the PI in a regression model. The Probabilistic Index Model (PIM) is such a regression model. For didactic reasons, we start by introducing PIMs for a univariate covariate. These univariate PIMs can then be extended to multiple covariates.

One Covariate

Consider a sample of n identically and independently distributed (i.i.d.) observations (Y_i, X_i). A PIM models the conditional PI directly as a function of the covariates. More specifically, a PIM is given by

$$\mathrm{P}\,(Y_i < Y_j | \, X_i,\, X_j) = m(\,X_i,\, X_j;\, \beta). \qquad (3)$$

Here $m(\cdot)$ is a user-specified function and β is the regression parameter that we want to estimate. As will be explained in the next section, the following choice of $m(\cdot)$ will be convenient for a variety of applications:

$$m\big(X_i,\, X_j;\, \beta\big) = \; g^{-1}[(X_j - X_i)\beta], \qquad (4)$$

where $g^{-1}(\cdot)$ denotes an inverse-link function mapping the real line on the unit interval. Examples include the logit with $g(x) = \mathrm{logit}(x) = \; \log[x/(1-x)]$ and $g^{-1}(x) = \mathrm{expit}(x) = \exp(\mathrm{x})\,/[1 + \exp(\mathrm{x})]$, or the probit with $g(x) = \Phi^{-1}(x)$ and $g^{-1}(x) = \Phi(x)$.

The interpretation of the regression coefficient can be obtained by comparing two individuals with covariates that differ by one unit: $X_i = x$ and $X_j = x + 1$. We substitute these values in Equation 4 and then replace this expression in Equation 3 to obtain:

$$g^{-1}(\beta) = \mathrm{P}(Y_i < Y_j | \, X_i = x,\, X_j = x + 1). \qquad (5)$$

Hence $g^{-1}(\beta)$ gives the probability that a randomly selected subject with covariate value x will have a lower outcome as compared to a randomly selected subject with covariate value $x + 1$. When X is binary, then the two sample PI P $\left(Y_i < Y_j \mid X_i = 0, X_j = 1\right)$ arises from Equation 5 by setting $x = 0$. The advantage of modeling the PI directly as a function of covariates is that Equation 3 can also be used when X is continuous. Equation 5 then gives the PI when X is increased by one unit. This is similar to the interpretation of a conventional linear regression model. When $\mathrm{E}(Y_i|X_i) = \alpha_0 + \alpha X_i$, then

$$\alpha = \mathrm{E}\left(Y_j \middle| X_j = x + 1\right) - \mathrm{E}(Y_i|X_i = x) = \mathrm{E}\left(Y_j - Y_i \middle| X_i = x, X_j = x + 1\right).$$

Whereas a PIM quantifies the effect in terms of the probability of an ordering between two outcomes, the linear regression models quantifies the effects in terms of the expected differences between two outcomes.

When X is binary, we know that we can estimate the PI via Equation 2. How can we estimate the PI when X is continuous? From Equation 5 we see that we can estimate the PI via $g^{-1}(\hat{\beta})$, where $\hat{\beta}$ denotes an estimator of β. This brings us to the question: How can we estimate β? The solution lies in rewriting the PI as an expectation. Recall that I $(Y_i < Y_j)$ is 1 if the outcome of subject i is lower than the outcome of subject j, and is 0 otherwise (see Equation 2). In the supplementary material we show that

$$\mathrm{P}\ (Y_i < Y_j|\ X_i, X_j) = \mathrm{E}\big(\ \mathrm{I}\ (Y_i < Y_j)|\ X_i, X_j\big)\ . \tag{6}$$

The probability statement on the left hand side of Equation 6 can thus be written as an expectation of transformed outcome I $(Y_i < Y_j)$ that takes on the value 0 or 1. We introduce the compact notation $I_{ij} = \mathrm{I}(Y_i < Y_j)$ and $X_{ij} = X_j - X_i$ and we consider the logit link to make the formulation more concrete. Combining Equations 3, 4 and 6 gives

$$\mathrm{E}\left(I_{ij} \middle| X_{ij}\right) = \mathrm{expit}(X_{ij}\beta).$$

This is exactly a logistic regression model applied to the transformed binary outcomes I_{ij}, and the transformed predictors $X_{ij}, (i, j = 1, \dots, n)$. Thas et al. (2012) show that this strategy results in an asymptotically normal and consistent estimator of β. Hence, fitting a PIM to the data (Y_i, X_i) is equivalent to fitting a binary regression model to the transformed data (I_{ij}, X_{ij}). By rewriting a PIM as a binary regression model, we can use existing software to fit PIMs in practice. More formally, the estimator $\hat{\beta}$ for β in Equation 3 is obtained by solving the estimating equations

$$\sum_{i=1}^{n}\sum_{j=1}^{n} A\left(X_i, X_j; \beta\right)\left[I_{ij} - m\left(X_i, X_j; \beta\right)\right] = 0,$$
$$A\left(X_i, X_j; \beta\right) = \frac{\frac{\partial m(X_i, X_j; \beta)}{\partial \beta}}{m(X_i, X_j; \beta)[1 - m(X_i, X_j; \beta)]}\,. \tag{7}$$

The standard errors obtained by fitting a binary regression model to the transformed data cannot be used for inference. Despite the data $(Y_i,\ X_i)$ being mutually independent, the transformed data $(I_{ij},\ X_{ij})$ are no longer mutually independent. To illustrate this, consider two transformed outcomes $I_{ij} = \mathrm{I}(Y_i < Y_j)$ and $I_{ik} = \mathrm{I}(Y_i < Y_k)$. Both binary outcomes share Y_i, making them no longer independent. The transformed outcomes I_{ij} have a correlation structure which is different from the typical block correlation structure in multilevel or longitudinal data (Thas et al, 2012). Thas et al. (2012) provide a consistent sandwich estimator for the standard errors that takes the correlation into account. These standard errors are implemented in the package **pim** (Meys et al., 2017) of the R programming language (R Core Team, 2017).

Note that rewriting a PIM as a binary regression model has implications for its computational complexity: Whereas the original sample is of size n, the transformed sample $(I_{ij},\ X_{ij})$ has n^2 elements because we compare all pairs of subjects i and j. Thas et al. (2012) show that it sufficient to only consider the pairs of subjects for which $i < j$, resulting in $n(n-1)/2$ comparisons.

In summary, the theory provided by Thas et al. (2012) and implemented in the **pim** package, provides a consistent estimator for β that is asymptotically normal and a consistent estimator for its standard error $(SE_{\hat{\beta}}\ )$.

Confidence intervals for β are given by $(\hat{\beta}\ - z_{\alpha/2}SE_{\hat{\beta}}, \hat{\beta}\ + z_{\alpha/2}SE_{\hat{\beta}})$, with $z_{\alpha/2}$ the quantile so that $\Phi(z_{\alpha/2}) = 1 -\ \alpha/2$. Because $g^{-1}(\cdot)$ is strictly increasing, a confidence interval for $g^{-1}(\beta)$ is given by $(g^{-1}[\hat{\beta}\ - z_{\alpha/2}SE_{\hat{\beta}}], g^{-1}[\hat{\beta}\ + z_{\alpha/2}SE_{\hat{\beta}}])$, see the supplementary material. Hypothesis tests for $H_0: \beta = \beta_0$ are obtained by constructing the test statistic $(\hat{\beta} - \beta_0)/SE_{\hat{\beta}}$ which has an asymptotic standard normal distribution under the null hypothesis.

Because these confidence intervals and hypothesis tests can perform poorly in small samples, mainly due to the small sample bias of the sandwich estimator, Amorim et al. (2017) proposed a bias-reduced version of the bootstrap and adjusted jackknife empirical likelihood that lead to drastic improvements in small sample inference for PIMs. Discussing these estimators, however, falls beyond the scope of this article.

Despite the close connection with logistic regression, it is important to mention that the link-function of a PIM plays a different role than the link function in a generalized linear model. In logistic regression, we model the probability of a 'success' as a function of the covariate via $\mathrm{P}(\mathrm{succes}|X_i) = \mathrm{expit}(\gamma_0 + \gamma X_i)$. From this model, we then derive an interpretation of γ via

$$\exp(\gamma) = \frac{\mathrm{odds}(\mathrm{succes}|X_i=x+1)}{\mathrm{odds}(\mathrm{succes}|X_i=x)}, \qquad \mathrm{odds}(\mathrm{success}|X_i) = \frac{\mathrm{P}(\mathrm{succes}|X_i)}{1-\mathrm{P}(\mathrm{succes}|X_i)}.$$

Here the choice of link function is crucial in obtaining this odds ratio interpretation. When a different link function is used, e.g. the probit, then $\exp(\gamma)$ has no interpretation in terms of an odds ratio. For a PIM, this is different. From Equation 5, we see that we can always transform β via $g^{-1}(\cdot)$ to get an interpretation in terms of the PI. This holds for all link functions. This is a consequence of the fact that a PIM models an effect size directly, whereas in logistic regression the effect sizes are derived from modeling the probability of success.

Multiple Covariates

The rationale of the previous section can be adopted to a multivariable context. Let $\boldsymbol{X}$ denote the *p*-dimensional vector of covariates associated with *Y*, then a PIM is given by

$$\mathrm{P}\,(Y_i < Y_j |\, \boldsymbol{X}_i, \boldsymbol{X}_j) = g^{-1}[\left(\boldsymbol{X}_j - \boldsymbol{X}_i\right)^T \boldsymbol{\beta}], \tag{8}$$

with $\boldsymbol{\beta}$ the *p*-dimensional vector of interest. To estimate $\boldsymbol{\beta}$ we transform the outcomes as before to I_{ij} and the vector of covariates to $\boldsymbol{X}_{ij} = \boldsymbol{X}_j - \boldsymbol{X}_i$. Similar to how logistic regression can deal with multiple covariates, so too can the PIM (Thas et al, 2012). The regression coefficient can be estimated by solving the estimating equations (7) where the function $m(\cdot)$ is replace by the right-hand side of Equation 8.

To illustrate the interpretation, we consider a bivariate regressor: $\boldsymbol{X}^T = (Z_1, Z_2)$ with $\boldsymbol{\beta}^T = (\beta_1, \beta_2)$. We consider two subjects, i and j, with covariate patterns $(Z_{1i}, Z_{2i}) = (z_1, z_2)$ and

$(Z_{1j}, Z_{2j}) = (z_1 + 1, z_2)$, respectively. Both subjects have the same value of Z_2, while Z_1 differs by one unit. It follows that $\left(\boldsymbol{X}_j - \boldsymbol{X}_i\right)^T = (1, 0)$ so that from the Equation (8) we obtain the following interpretation of the regression coefficient:

$$\mathrm{P}\left(Y_i < Y_j \middle| Z_{1i} = z_1, Z_{1j} = \mathrm{z}_1 + 1, Z_{2i} = Z_{2j}\right) = g^{-1}(\beta_1),$$

i.e., $g^{-1}(\beta_1)$ gives the probability that a randomly selected subject with covariate value z_1 for Z_1 will have a lower outcome as compared to a randomly selected subject with a covariate value that is one unit higher, where Z_2 is the same for both subjects. Similar as for the univariate covariate, we can estimate this PI by plugging in an estimate for β_1.

Notice that the PIM of Equation 8 has no intercept. This can be explained as follows: If $\boldsymbol{X}_j - \boldsymbol{X}_i = \boldsymbol{0}$, then $\boldsymbol{X}_i = \boldsymbol{X}_j$ so that Equation 8 reduces to

$$\mathrm{P}\left(Y_i < Y_j \middle| \boldsymbol{X}_i = \boldsymbol{X}_j\right) = g^{-1}(0).$$

For both the logit and probit link it follows that $g^{-1}(0) = .5$. The model thus implies that the PI equals 50% when the covariates of both subjects are the same. Indeed, when $\boldsymbol{X}_i = \boldsymbol{X}_j$ then Y_i and Y_j have the same conditional distribution so that $\mathrm{P}\left(Y_i < Y_j \middle| \boldsymbol{X}_i = \boldsymbol{X}_j\right) = 0.5$ has to hold. Setting the intercept to zero thus automatically satisfies this restriction.

Dealing with Ties

The definition of the PI can be extended to $\mathrm{P}\left(Y_i < Y_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right) + \frac{1}{2}\mathrm{P}\left(Y_i = Y_j \middle| \boldsymbol{X}_i, \boldsymbol{X}_j\right)$ to account for ties. The PIM estimation theory can now be adopted by considering the transformed outcomes $\mathrm{I}(Y_i < Y_j)$ + 0.5 $\mathrm{I}(Y_i = Y_j)$. The remainder of the estimation theory is unaffected by these ties. We refer to Thas et al. (2012) for more details.

Estimating PIMs in Practice

In R, PIMs can be fitted using the package **pim**. The online supplementary material includes all R code to reproduce the results presented in this article. For other statistical packages, no ready-to-use software is yet available. However, logistic and probit regression routines can be used to obtain point estimators for the regression coefficients of a PIM. Bootstrap procedures can then be used for obtaining interval estimators and p-values. This can be achieved as follows.

1. For each pair of subjects i and j create transformed outcomes $I_{ij} = \mathrm{I}(Y_i < Y_j) + 0.5\mathrm{I}(Y_i = Y_j)$ and transformed predictors $\boldsymbol{X}_{ij} = \boldsymbol{X}_j - \boldsymbol{X}_i$. If the predictors include categorical variables, we assume that $\boldsymbol{X}$ is numerically coded (e.g. using a dummy coding). When subject i is compared with subject j, it is not necessary anymore to compare subject j with i. If the original sample size is n, then the transformed data will have size $n(n-1)/2$. To facilitate creating this transformed data, a shiny app has been developed at datapp.ugent.be/shiny/trans4pim/.
2. For a PIM with logit link, fit a logistic regression model with outcome I_{ij} and predictor $\boldsymbol{X}_{ij}$. The logistic estimates are now exactly the estimates proposed by Thas et al. (2012). For a PIM with probit link, probit regression should be considered.
3. The standard errors and the p-values from a logistic (or probit) regression routine cannot be used because they do not take into account the correlation structure of the transformed outcomes. Bootstrapping can resolve this as follows: Sample with replacement B bootstrap samples of size n from the original data (not the transformed data) and repeat step 1 and 2 for each bootstrap sample. Compute the standard error as the standard deviation of the B bootstrap samples and use this standard error to compute p-values and to construct confidence intervals as explained at the beginning of this section.

Comparison with Other Methods

For a better understanding of PIMs, we study its relationship to several other methods. We start with the connection between the Wilcoxon–Mann–Whitney test and the parametric regression model with normal errors. These relationships are then extended to a semiparametric context. To make the comparison more concrete, we illustrate each connection on a case study. We consider data from a clinical trial where a computerized, interactive cognitive behavioral therapy for patients with depression is evaluated. The original study is reported in Proudfoot et al. (2003), and the data are available in the R package of Hothorn and Everitt (2017a,b). Patients with depression were recruited in primary care and were randomized over two treatments: an innovative treatment or a conventional treatment. The conventional treatment (coded as $X = 1$ with 36 patients) consisted of face-to-face cognitive behavioral therapy, while the innovative treatment consisted of an interactive computerized program called Beat the Blues™ (coded as $X = 0$, with 37 patients) replacing the face-to-face counselling. The outcome of interest Y is the Beck Depression Inventory (BDI) II score after three months. Figure 4 shows boxplots of the depression score for both treatments. For the ease of exposition, we consider a univariate predictor in this section. We fit the PIM

$$P\big(Y_i < Y_j \mid X_i, X_j \big) = \Phi\big[\beta\big(X_j - X_i\big)\big]. \tag{9}$$

The estimated regression coefficient equals $\hat{\beta} = 0.357$ with standard error 0.17 and 95% confidence interval [0.02, 0.69] ($p = .036$). We first compare these results with those of the WMW test, and then with the results of a normal linear regression model.

Wilcoxon-Mann-Whitney Test

The WMW test shows marginal evidence against the null hypothesis $H_0\colon F_{IT} = F_{CT}$ in favor of the alternative $H_A\colon P(Y_{IT} < Y_{CT}) \neq .5$ ($p = .041$). Here F denotes the depression score distribution function. The WMW statistic from Equation 2 equals $MW = .639$: The probability that a patient receiving the innovative treatment will have a lower depression score as compared to a patient receiving the conventional treatment is estimated as 63.9%. De Neve and Thas (2015) derived the following explicit relationship between WMW statistic and the estimated PIM parameter: $\Phi(\widehat{\beta)} = MW$. Applied to the case study we have $\Phi(0.357) = .639$, confirming this relationship. Using a PIM in a two-sample study is thus similar to using the WMW test. When we look at the p-values, we notice that they are similar but not identical: .036 for the PIM and .041 for the WMW test. This can be explained as follows: Both p-values are derived from a standardized statistic. The standard error used by the WMW test (in absence of ties) equals $\sqrt{(n_{IT} + n_{CT} + 1)/12\, n_{IT}n_{CT}}$, and is only correct under $H_0\colon F_{IT} = F_{CT}$. The standard error estimator of the PIM, on the other hand, is also consistent when this null hypothesis does not hold. This explains the difference in p-values: They rely on standardized statistics that use different estimators of the standard error. The standard error provided by the PIM has three consequences:

1. It allows testing $H_0\colon \beta = 0$ versus $H_A\colon \beta \neq 0$. From Equation 9 we see that $P(Y_{IT} < Y_{CT}) = P\big(Y_i < Y_j \mid X_i = 0, X_j = 1 \big) = \Phi[\beta(1-0)] = \Phi(\beta)$, so the PIM tests $H_0\colon P(Y_{IT} < Y_{CT}) =$

$\Phi(0) = .5$ versus $H_A: P(Y_{IT} < Y_{CT}) \neq .5$. This null hypothesis is less restrictive than the null hypothesis of the WMW test in the sense that equal distributions imply that the PI equals .5, while the opposite does not necessarily hold. We refer to the supplementary material for more details.

2. Because the standard error has to be estimated, the small sample performance (e.g. control of the Type I error) of the test based on the PIM will be not as good as that of the WMW test because the standard error under the null of equal distribution does not involve parameters that need to be estimated.
3. It allows for the construction of confidence intervals for β, and hence $\Phi(\beta) = P(Y_{IT} < Y_{CT})$. We refer to the previous section for details on how the confidence interval can be constructed.

This last point is the most important implication: Using PIMs allows going beyond null hypothesis significance testing by providing interval estimators for the PI effect measure. This might lead to a better understanding of the data under study.

Similar connections can be established for other rank tests, such as the Kruskal–Wallis rank test and the Friedman rank test; we refer to De Neve and Thas (2015) for details.

The Normal Linear Model

The linear regression model with normal errors is given by

$$Y_i = \alpha_0 + \alpha_1 X_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2). \tag{10}$$

In the supplemental material we show that this normal linear model implies

$$\mathrm{P}\left(Y_i < Y_j \middle| X_i, X_j\right) = \Phi\left(\frac{\alpha_1(X_j - X_i)}{\sqrt{2}\sigma}\right). \tag{11}$$

From this relationship we recognize the form of Equation 9 with $\beta = \alpha_1/\sqrt{2}\sigma$. This derivation implies that the formulation of Equation 9 arises naturally when a normal linear model holds.

To make this relationship more concrete, we fit the linear model to the depression case study. With BDI score as outcome, the residuals of the linear regression model indicated a significant deviation from normality (Shapiro-Wilk $p = .0008$). Transforming the BDI score by taking the square root resolved the deviation from normality (Shapiro-Wilk $p = .76$) and we therefore continue with this transformed outcome: $Y = \sqrt{BDI}$. The estimated slope is $\hat{\alpha}_1 = 0.8$ (*SE* = 0.36, p = .031, 95% CI [0.08,1.51]), indicating that the mean square root BDI score is an estimated 0.8 units lower for the innovative treatment as compared to the conventional treatment. The maximum likelihood estimate of the standard deviation is $\hat{\sigma} = 1.52$. From Equation 11 it follows that

$$P(Y_{IT} < Y_{CT}) = \mathrm{P}\left(Y_i < Y_j \middle| X_i = 0, X_j = 1\right) = \Phi\left(\frac{\alpha_1}{\sqrt{2}\sigma}\right),$$

and upon plugging in $\widehat{\alpha_1}$ and $\hat{\sigma}$ the PI is estimated as 64.4% (using the unrounded estimates). This estimate is close to the estimate 63.9% obtained from the PIM of Equation 9. Recall, however, that the PIM is fitted to the BDI score and not to its square root. Both estimates are still close because the PIM estimator of β is not affected by this square root transformation. This can be seen by looking at Equation 7, taking into account that $I(Y_i < Y_j) = I[h(Y_i) < h(Y_j)]$ for every order preserving function $h(\cdot)$ (including the square root). Consequently, transforming the outcome via any strictly increasing function has no impact on the results of a PIM: The estimates, p-values, and confidence intervals will not change.

So far, we have discussed two ways to estimate $P(Y_{IT} < Y_{CT})$. The first approach uses $\Phi(\hat{\beta})$ where $\hat{\beta}$ is the estimator of the PIM from Equation 9. This is exactly the same as using the nonparametric Mann-Whitney statistic of Equation 2. The second approach consists of plugging in the linear regression estimators of α_1 and σ in Equation 11. For brevity, we refer to the first estimator as the nonparametric estimator, and the second as the parametric estimator. Below we list the differences between the two approaches.

1. *Bias and consistency*. The parametric estimator is only valid when the normal linear model holds and will be biased because $E[\,\Phi\left(\frac{\hat{\alpha}_1}{\sqrt{2}\hat{\sigma}}\right)] \neq \Phi(\frac{\alpha_1}{\sqrt{2}\sigma})$. When the errors in Equation 10 are not normally distributed the parametric estimator will in general not be a consistent estimator for the PI. The nonparametric estimator, on the other hand, does not require any distributional assumptions: It is an unbiased and consistent estimator of the PI for every distribution. It is also the estimator with the smallest standard error among all unbiased estimators (Lehmann, 1951).
2. *Robustness.* The nonparametric estimator is robust against outlier because it only involves the ordering of the outcomes, i.e. $I(Y_i < Y_j)$, and not their actual values. The parametric estimator, on the other hand, will be sensitive to outliers because it relies on the linear regression estimators.

3. *Transformations*. The nonparametric estimator is not affected by an order preserving transformation of the outcome. This does not hold for the normal least squares estimator.

From Parametric to Semiparametric Regression

When we leave the two-sample design and let X be a continuous covariate, we cannot use the WMW test anymore. On the other hand, the PIM of Equation 9 can still be used, because the PIM estimation theory holds for any type of covariate (binary, continuous, multivariable). From Equation 11 we can justify this probit PIM formulation: If the normal linear model from Equation 10 holds, Equation 9 follows automatically from Equation 11. This might give the impression that the probit PIM is a parametric model which is only valid when the parametric normal linear model holds. This, however, is untrue. PIMs are substantially less restrictive than the normal model. This can be seen as follows: The left-hand side of Equation 9 is the same for any order preserving transformation $h(\cdot)$ of the outcome:

$$\mathrm{P}\left(h(Y_i) < h(Y_j) \middle| X_i, X_{\boldsymbol{j}}\right) = \mathrm{P}\left(Y_i < Y_j \middle| X_i, X_{\boldsymbol{j}}\right).$$

This implies that, from the moment we can find *any* order preserving transformation of the outcome for which the normal linear model holds, the relationship from Equation 11 holds, and the PIM with probit link from Equation 9 is justified. We can write this down more formally. We assume that the following model holds

$$h(Y_i) = \alpha_0 + \alpha_1 X_i + \epsilon_i \text{ , } \epsilon_i \sim N(0, \sigma^2) \tag{12}$$

where $h(\cdot)$ denotes an *unknown* strict increasing function. This model is semiparametric because $h(\cdot)$ is left unspecified (the nonparametric component) and ϵ_i has a fully specified distribution (the parametric component). The model in Equation 12 is known as the semiparametric linear transformation model (Cheng, Wei, & Ying, 1995; Zeng & Lin, 2007). In the supplementary material we show that Equation 12 implies Equation 11.

What is the most important conclusion from this? A PIM with probit link is justified from the moment that there exists an unknown strictly increasing function $h(\cdot)$ so that the regression model with transformed outcome $h(Y)$ has a normally distributed error. This function $h(\cdot)$ can be as simple as the identity function or the square root, but it can also be more complicated, such as one or two-parameter Box-Cox transformations. For a PIM we do not need to estimate this function, it merely serves as theoretical argument to demonstrate its semiparametric nature and to justify the functional form of the PIM as described in Equation 9. As we will demonstrate in the next section, the relationship between Equations 9 and 12 can be used to assess the goodness-of-fit of a PIM.

So far, we have focused on the PIM with probit link. This leads to the question: Does it make sense to use a PIM with logit link? The answer is yes, we can think of data generating models that result in a logit PIM. In the supplementary material we show that

$$h(Y_i) = \alpha_0 + \alpha_1 X_i + \epsilon_i, \qquad \epsilon_i \sim F_{EV}(e) = 1 - \exp[-\exp(e)], \tag{13}$$

where $F_{EV}(\cdot)$denotes the extreme value distribution, implies

$$\mathrm{P}\left(Y_i < Y_j \middle| X_i, X_{\boldsymbol{j}}\right) = \mathrm{expit}[\alpha_1(X_j - X_i)]\textbf{.}$$

Consequently, the semiparametric transformation model assuming an extreme value error implies that a PIM with logit link (recall that expit is the inverse of the logit) holds. The model of Equation 13 is also known as the Cox proportional hazards model (Cox, 1972, Cheng et al, 1995), a model that is extensively used in biostatistics.

The choice of link function, logit or probit, will have little impact on the estimated PIs in practice, because $\Phi\left(\frac{x}{1.7}\right) \approx \mathrm{expit}(x)$, implying that the estimated coefficients from a probit PIM will approximately differ by a factor 1.7 of the estimated coefficients of a logit PIM. We refer to the supplementary material for more details.

Goodness-of-fit

Because PIMs are semiparametric, inference is only valid when the proposed model is consistent with the underlying data-generating model. Therefore, it is important to assess the goodness-of-fit (GOF) of a PIM. De Neve et al. (2013a) developed a formal GOF-test together with GOF-plots for PIMs. These methods rely on nonparametric smoothers and require large datasets when there are multiple predictors. This makes them not useful in many practical situations. We will therefore address the GOF differently. More specially, we will exploit the connection with the semiparametric linear transformation models of the previous section. We consider the following procedure:

1. Fit the PIM

$$P\left(Y_i < Y_j \mid \boldsymbol{X}_i, \boldsymbol{X}_j \right) = \Phi\left[\left(\boldsymbol{X}_j - \boldsymbol{X}_i \right)^T \boldsymbol{\beta} \right].$$

2. Check the assumptions of the linear model (linearity of the model and constant variance and normality of the errors). If they are fulfilled, go to step 3. If some of the assumptions are violated, go to step 4.
3. Due to the connection with the normal linear model, the PIM can be used for further analysis.
4. Perform a Box–Cox transformation on the linear regression model.
5. Check the assumptions of the linear model applied to the Box–Cox transformed outcome (linearity of the model, constant variance, and normality of the errors). If they are fulfilled, go to step 6. If some of the assumptions are violated, go to step 7.
6. Due to the connection with semiparametric linear transformation model with normal error, the PIM can be used for further analysis.
7. If the linearity of the model of the Box-Cox transformed model is violated, try to include quadratic terms or splines in the PIM. If the residuals show a systematic deviation from normality and they show an approximate extreme value distribution, the probit link can be replace with the logit link.

The above procedure is heuristic and indicates that more research on assessing the adequacy of PIM is needed.

Full Illustration

We illustrate the methodology on data from a cross-sectional lifespan investigation of deception in a large community sample. In the original study, reported in Debey, De Schryver, Logan, Suchotzki, and Verschuere (2015), the ‘skill’ to deceive was measured using the Sheffield lie test. This test is a reaction time (RT) deception task whereby participants are instructed to alternately lie or tell the truth on 120 simple yes/no questions. Because the truth response is, in general, assumed to be activated first, faster responses are expected for truth trials compared to lie trials. The outcome of interest is the difference in the median reaction time between lie trials and truth trials. This difference is referred to as the RT lie-effect, and larger values indicate a larger lie-effect. A larger lie-effect for a participant can be interpreted as that participant being less skilled at lying.

For the current illustration, the median is used to aggregate the reaction time per participant, in order to reduce the impact of outlying values. After excluding cases with missing values, and participants with a negative lie-effect, 831 participants remain in the present analysis. We consider the following variables: the RT lie-effect (ranging from 2 ms to 2819 ms, with a mean of 253 ms and a median of 207 ms), gender (42% male, 58% female), and age (ranging from 6 to 76 years, with a mean age of 28 and a median age of 20). It is of interest for this illustration to study the association between the lie-effect and age while accounting for gender. Panel a of Figure 5 displays the RT lie-effect according to age group, where groups are formed as in Debey et al. (2015). The plot demonstrates a non-linear effect of age. Panel b of Figure 5 shows a scatterplot with the fit of a linear regression with age included as a cubic polynomial.

We fit the PIM with the RT difference as outcome (Y), a linear, a quadratic, and cubic effect of age (variable A) and a dummy variable for gender (variable G which equals 1 for females). The vector of covariates then becomes $\boldsymbol{X}_i^T = (A_i, A_i^2, A_i^3, G_i)$, and plugging this in Equation 8, upon using the probit link, gives

$$P(Y_i < Y_j) = \Phi\left[(A_j - A_i)\beta_1 + (A_j^2 - A_i^2)\beta_2 + (A_j^3 - A_i^3)\beta_3 + (G_j - G_i)\beta_4\right], \qquad (14)$$

where we have omitted the conditioning statement in the probability for notational convenience. The probit link is considered after a careful exploration of the goodness-of-fit of the PIM. We refer to the R Markdown supplemental material for more details.

The estimates are $\widehat{\beta_1} = -0.084$ (*SE* = 0.021, $p <$ .001), $\widehat{\beta_2} = 0.003$ (*SE* = 0.0006, $p <$ .001), $\widehat{\beta_3} = -0.00003$ (*SE* = 0.00001, $p <$ .001), and $\widehat{\beta_4} = 0.094$ (*SE* = 0.051, $p =$ .07). To illustrate the interpretation, we compare subjects of the same gender ($G_i = G_j$). Figure 6 visualizes the non-linear effect of age. The PI always refers to the outcomes of two subjects, (Y_i, Y_j), and we consider 3 choices for the age of subject i in Figure 6: (a) a randomly chosen subject with age $A_i = 20$ which we compare to randomly chosen subjects with ages varying from 6 to 76, (b) a randomly chosen subject with age $A_i = 40$ which we compare to randomly chosen subjects with ages varying from 6 to 76, and (c) a randomly chosen subject with age $A_i = 60$ which we compare to randomly chosen subjects with ages varying from 6 to 76. This plot is the result of plugging in these values in Equation 14. When we compare a subject of age 20 (dashed line) with any other subject, probabilities are estimated to be larger than 50% (note that if we compare two subjects of age 20, the probability of observing a smaller lie-effect equals 50%): It is more likely that a given twenty year old will have a smaller lie effect (i.e. a lower outcome) compared to subjects that are older or younger. If we compare this 20-year-old with someone of age 60, the estimated probability of observing a smaller lie-effect for the younger subject equals about 80%: It is more likely to observe a smaller lie effect at the age of 20.

Now let us compare a subject of age 40 with younger subjects. From the solid line in Figure 6, we can see that the estimated probabilities to observe smaller lie-effects are less than 50%: Compared to their younger counterparts, 40-year-old subjects seemed to be less skilled at deception. On the other hand, if we compare them with older subjects, the estimated probabilities are above the 50% line (ignoring the probabilities for subjects older than 72 years). Hence, compared to older people, a 40-year-old person will likely be more gifted with lie skills. Finally, estimated probabilities comparing a 60-year-old subject with other subjects are depicted by the dotted line. All estimated probabilities are less than 50%, indicating that it would be rather unlikely to observe a smaller lie-effect for a 60-year-old person compared to anyone else.

The PIM in Equation 14 further allows to estimate the effect of gender while controlling for age. When we compare men and women of the same age, we obtain $P\left(Y_i < Y_j \mid G_i = 0, G_j = 1, A_i = A_j\right) = \Phi(\beta_4)$, which is estimated by $\Phi(0.094) = 0.54$. Therefore, the estimated probability that men will have a smaller lie effect than women of the same age is 54%.

Discussion

This article serves as an introduction to PIMs: A class of regression models where the association between an outcome and a covariate can be expressed in terms of the PI. This is fundamentally different from (generalized) linear regression models where associations are expressed in terms of mean differences. We have devoted a substantial part of this tutorial to the discussion of the PI as an effect measure, because a good understanding of PIMs starts with a good understanding of the PI. Furthermore, a PIM can be seen as a flexible procedure to estimate the PI for a variety of designs and datatype. So when can it be appropriate to analyze your data with a PIM? We believe that PIMs are suitable when the PI is an appropriate effect measure to summarize the association between the covariate and the outcome. Examples where the PI can be useful include outcomes with skewed distributions, outcomes that contain outliers, and ordinal outcomes. PIMs might also be considered when the outcome requires a monotone transformation (e.g. Box-Cox transformations) before linear regression can be applied. These transformations can obscure the interpretation of the linear regression coefficient, while they do not affect the PIM. PIMs are semiparametric regression models and do not require strong distributional assumptions such as normality. This does not mean that they are free of assumptions: The functional form of model has to be correctly specified. If important predictors are missing, or if the link function is not supported by the data, the statistical inference of the PIMs will not be valid.

We want to stress that we do not see the PIM as a replacement of the linear regression model (or any other regression model). Instead, we consider it as a complementary, robust tool which the analyst can use to gain more insights into the psychological processes that generate the data. When several regression models are used in a single analysis, family-wise type I errors can be inflated due to multiple testing, and this should be accounted for. The decision to use a PIM should be made before the data are analyzed, or on features of the outcome (e.g. ordinal, skewed or outliers), and should not be based on the p-values it generates.

Via this tutorial, we hope to stimulate the community to apply and investigate these models in the behavioral sciences. Despite several publications on PIMs and availability of the R package **pim**, more research has to be conducted to make these models applicable for many practical situations. Challenges include: (a) the extension of PIMs to multilevel data, (b) embedding latent variables in PIMs, (c) extending goodness-of-fit methods, (d) providing R^2-like measures, and (e) developing software for statistical packages other than R.

References

Acion, L., Peterson, J., Temple, S., and Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, 25:591–602.

Amorim, G., Thas, O., Vermeulen, K., Vansteelandt, S., and De Neve, J. (2018). Small sample inference for probabilistic index models. *Computational Statistics & Data Analysis*, 121:137-148.

Brumback, L. C., Pepe, M. S., and Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine*, 25(4):575–590.

Cheng, S.,Wei, L., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82(4):835– 845.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cox, D. (1972). Regression models and life tables (with Discussion). Journal of the Royal *Statistical Society. Series B*, 34:187–220.

Debey, E., De Schryver, M., Logan, G. D., Suchotzki, K., & Verschuere, B. (2015). From junior to senior Pinocchio: A cross-sectional lifespan investigation of deception. *Acta Psychologica*, *160*, 58-68.

De Neve, J., Meys, J., Ottoy, J.-P., Clement, L., and Thas, O. (2014). UnifiedWMWqPCR: the unified Wilcoxon–Mann–Whitney test for analyzing RT-qPCR data in R. *Bioinformatics*, 30(17):2494–2495.

De Neve, J. and Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, 110(511):1276–1283.

De Neve, J., Thas, O., and Ottoy, J.-P. (2013a). Goodness-of-fit methods for probabilistic index models. *Communications in Statistics - Theory and Methods*, 42(7):1193–1207.

De Neve, J., Thas, O., Ottoy, J.-P., and Clement, L. (2013b). An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data. *Statistical Applications in Genetics and Molecular Biology*, 12(3):333–346.

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.

Fay, M, P., Brittain, E. H., Shih, J. H., Follmann, D.A., and Gabriel, E.E. (2018). Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Statistics in Medicine*, doi: 10.1002/sim.7799.

Grissom, R. J. and Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, 6(2):135.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods* (Vol. 751). John Wiley & Sons.

Hothorn, T. and Everitt, B. S. (2017a). *A handbook of statistical analyses using R*. CRC press.

Hothorn, T. and Everitt, B. S. (2017b). HSAUR3: *A Handbook of Statistical Analyses Using R* (3rd Edition). R package version 1.0-6.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, pages 165–179.

Lehmann, E. L. (2004). *Elements of large-sample theory*. Springer Science & Business Media.

Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

Meys, J., De Neve, J., Sabbe, N., and Guimaraes de Castro Amorim, G. (2017). *pim: Fit Probabilistic Index Models*. R package version 2.0.1.

Morris, S. and DeShon, R. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-group designs. *Psychological Methods*, 7:105–125.

Proudfoot, J., Goldberg, D., Mann, A., Everitt, B., Marks, I., and Gray, J. (2003). Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice. *Psychological Medicine*, 33(02):217–227.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods,* 13(1):19.

Senn, S. (1997). Letter to the editor: Testing for individual and population equivalence based on the proportion of similar responses, by D. M. Rom and E. Hwang, Statistics in Medicine, 15,1489-1505 (1996). *Statistics in Medicine*, 16(11):1303–1305.

Senn, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects by L. Acion, J. Peterson, S. Temple and S. Arndt. *Statistics in Medicine*, 25:3944–3948.

Senn, S. (2011). U is for unease: Reasons for mistrusting overlap measures for reporting clinical trials. *Statistics in Biopharmaceutical Research*, 3:302–309.

Thas, O., De Neve, J., Clement, L., and Ottoy, J.P. (2012). Probabilistic index models (with Discussion). *Journal of the Royal Statistical Society - Series B*, 74:623–671.

Tian, L. (2008). Confidence intervals for P(Y1 > Y2) with normal outcomes in linear models. *Statistics in Medicine*, 27:4221–4237.

Van der Vaart, A. W. (1998). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Vermeulen, K., Thas, O., and Vansteelandt, S. (2015). Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment. *Statistics in Medicine,* 34(6):1012–1030.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.

Zeng, D. and Lin, D. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B,* 69(4):507–564.

Table 1

Asymptotic relative efficiency (ARE) of the *t*-test versus the WMW test

	Distribution								
	max(1-y^2,0)	Normal	Uniform	Logistic	t_3	Laplace	t_5	Exp	Cauchy
ARE	0.86	0.95	1	1.1	1.24	1.5	1.9	3	∞

Note. The ARE of the two-sample t-test relative to the WMW test for different distributions. A value larger than 1 implies that the WMW is (asymptotically) more efficient.

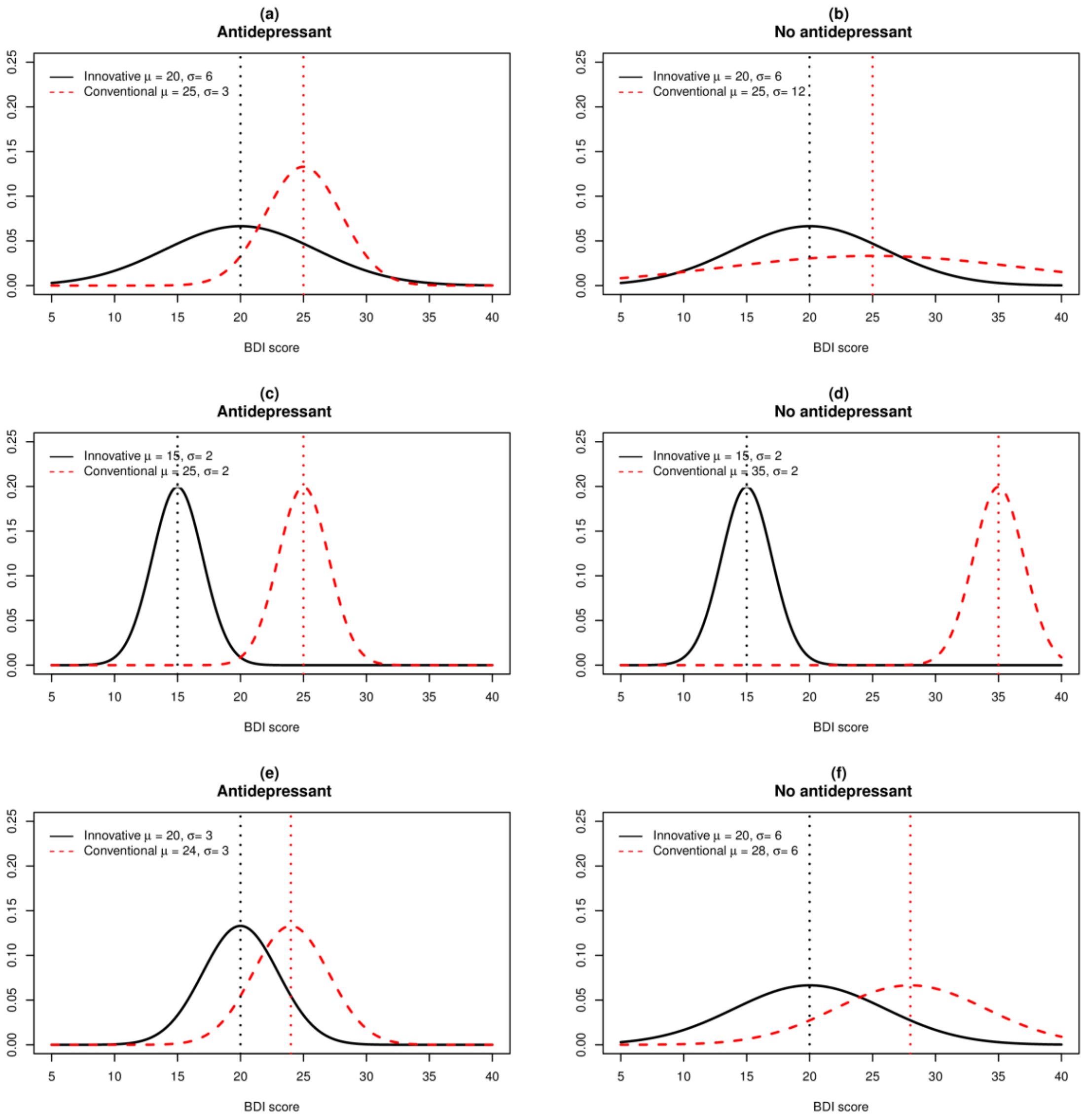

Figure 1. Artificial data where the innovative treatment outperforms the conventional treatment in terms of BDI scores for patients that receive antidepressants (panels a, c, e) or that do not receive antidepressants (panels b, d, f). All distributions considered are normal distributions and the mean and standard deviation are given for each group. Panels a and b: the standardized mean difference and the PI are larger for panel a, while the mean difference is the same for both panels. For panels c and d the mean difference and its standardized version is smaller for panel c, while the PI is approximately 1 for both panels. The mean difference in panel f is larger than the mean difference of panel e, while the standardized mean difference and the PI are equal for both panels.

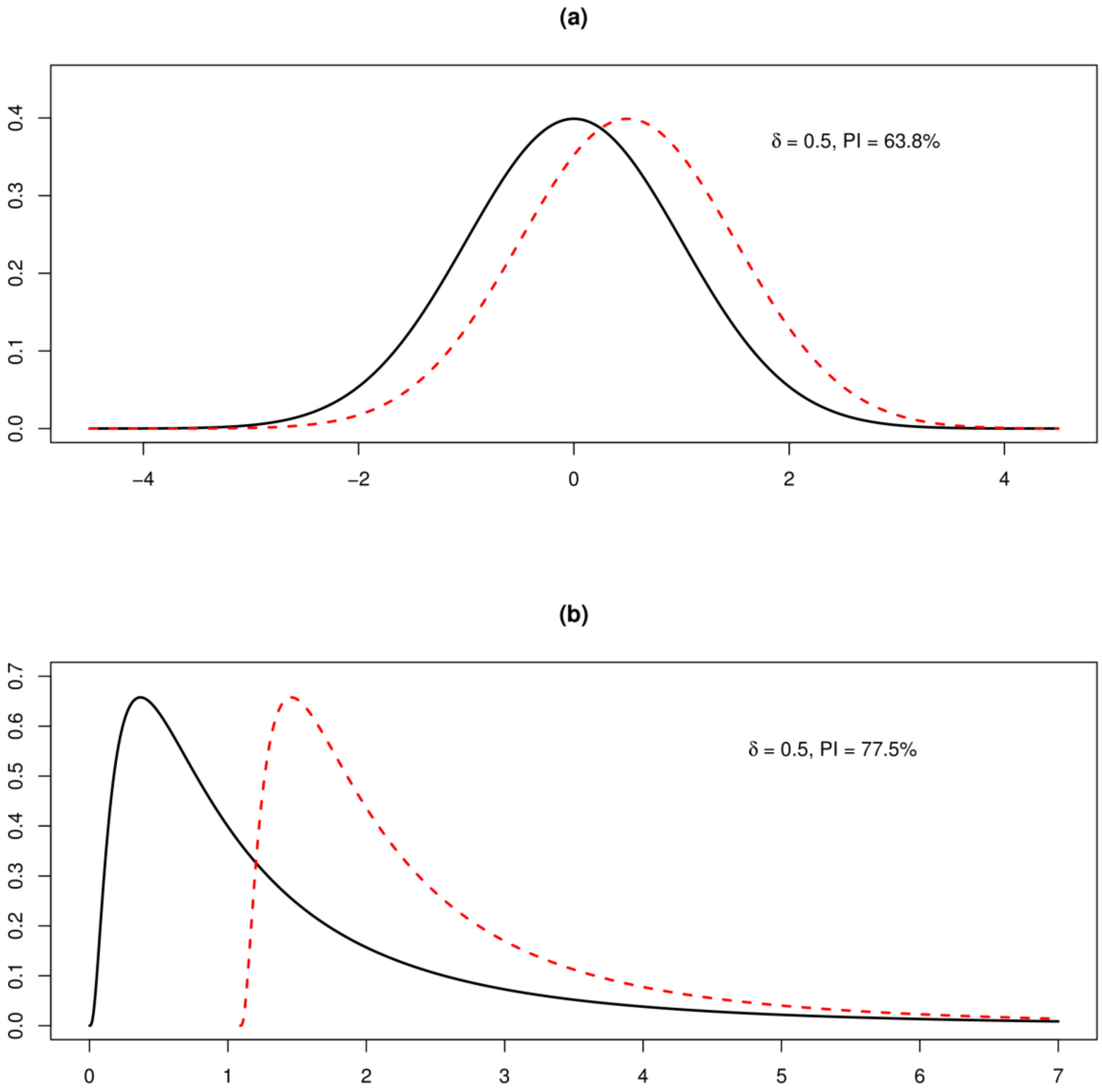

Figure 2. Two normal distributions (panel a) and two log normal distributions (panel b) for which the standardized mean difference is equal, but the PI's differ.

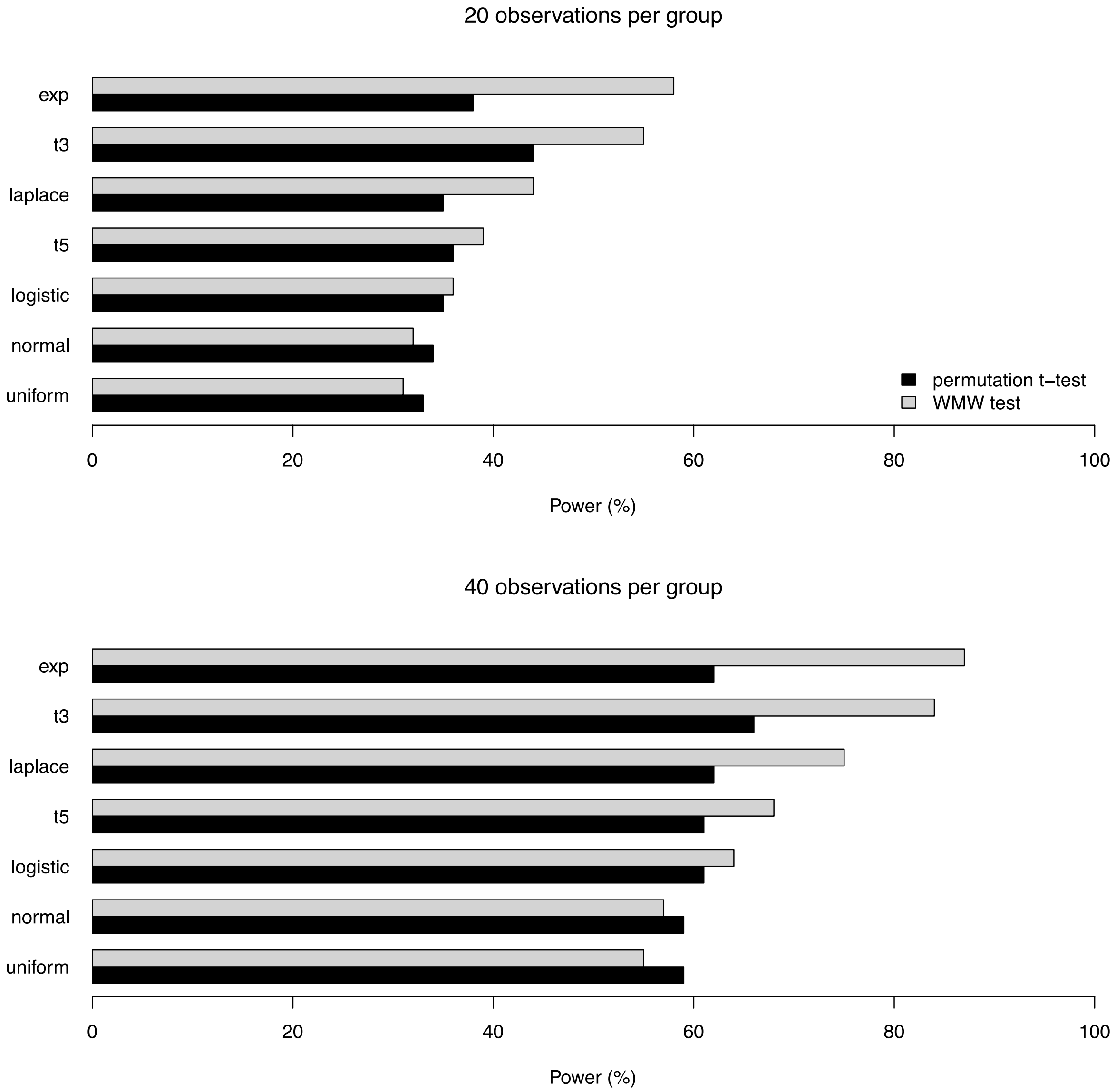

Figure 3. Power for a balanced two-sample design with 20 or 40 observations per group and for several choices of distributions. Both groups have the same distribution except for half standard deviation difference in location.

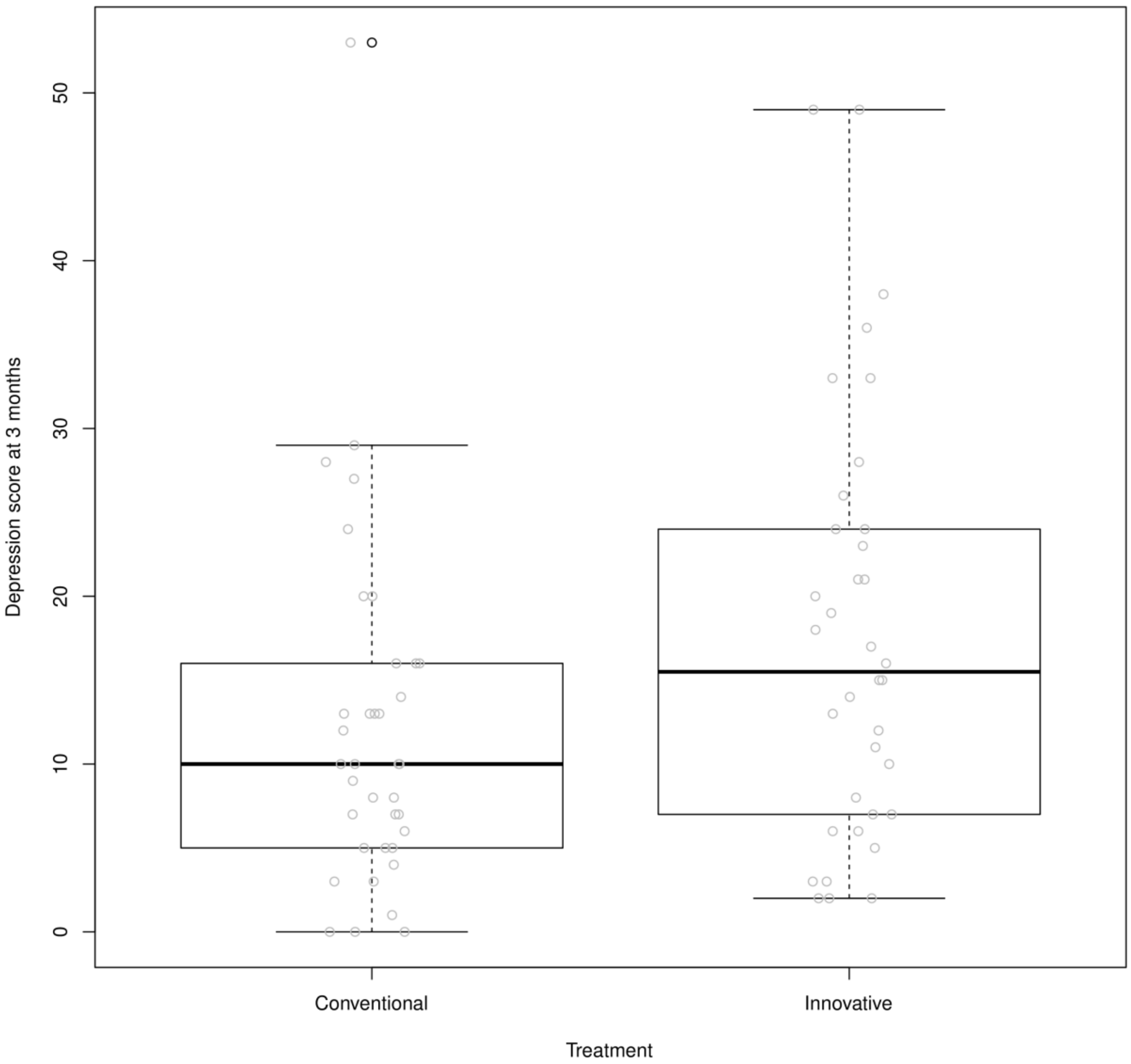

Figure 4. BDI scores at 3 months according to treatment group

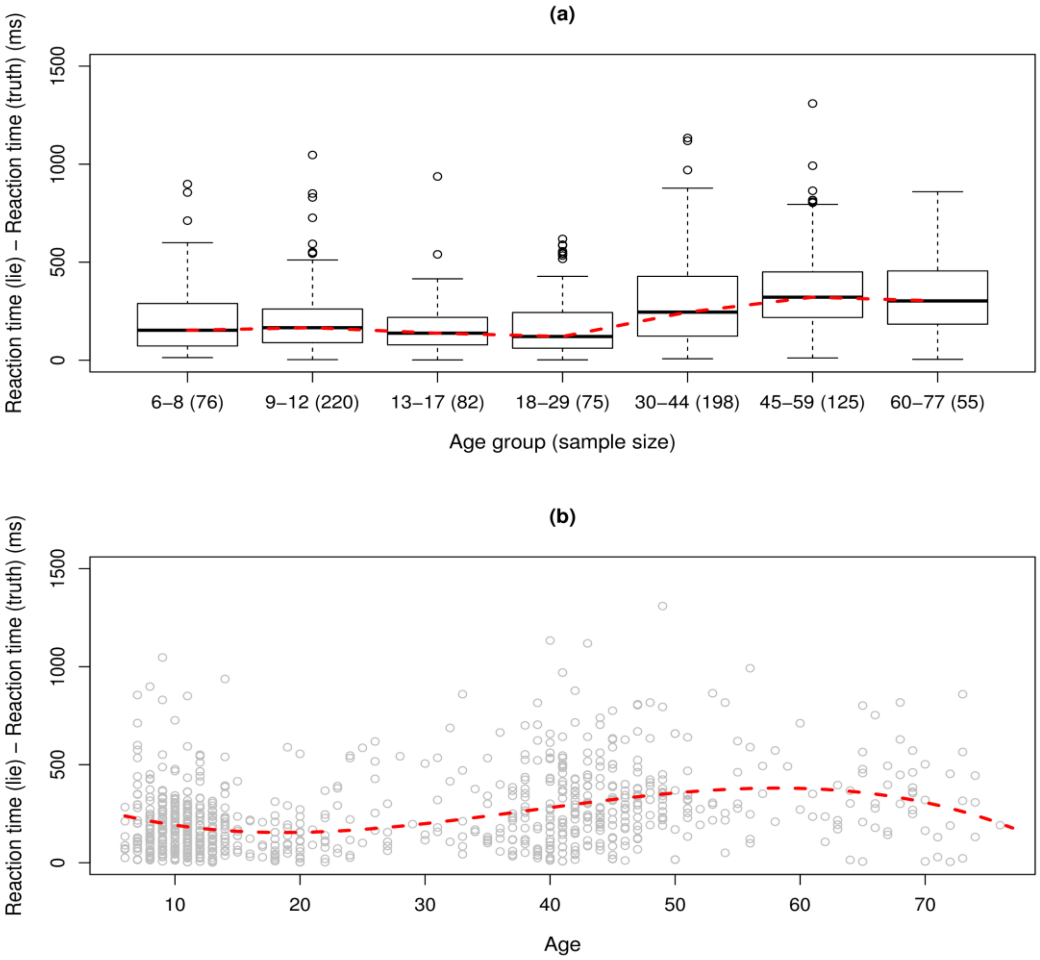

Figure 5. Difference in the median reaction time between lie trials and truth trials as a function of age category (panel a) and age (panel b).

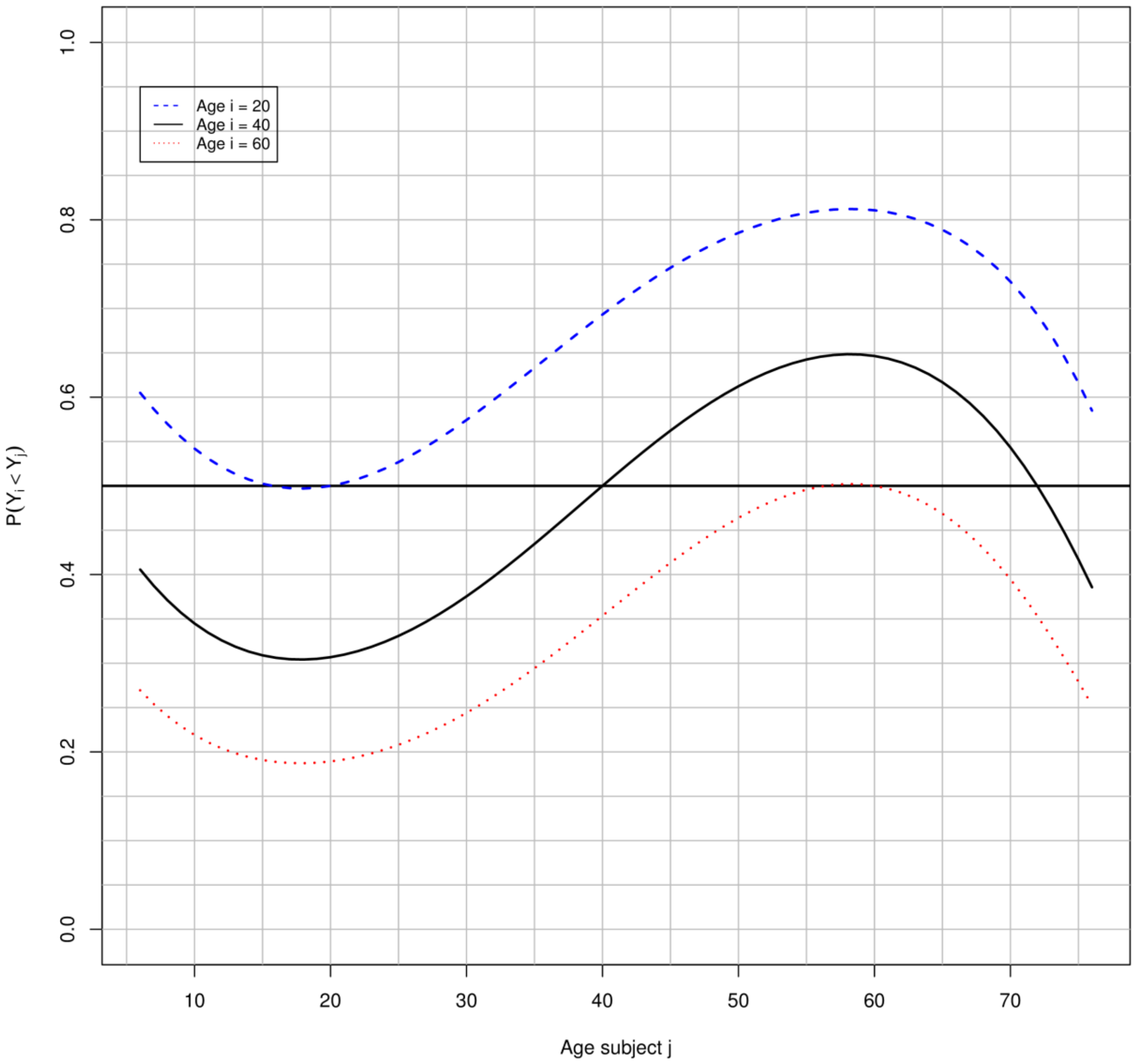

Figure 6. Estimated probabilities for a randomly chosen subject with age $A_i = 20$ compared to randomly chosen subjects with ages varying from 6 to 76 (dashed line), a randomly chosen subject with age $A_i = 40$ compared to randomly chosen subjects with ages varying from 6 to 76 (solid line), and a randomly chosen subject with age $A_i = 60$ compared to randomly chosen subjects with ages varying from 6 to 76 (dotted line).