pim: An R package for fitting probabilistic index models

Jan De Neve and Joris Meys

January 8, 2025

Contents

1 Introduction

In this document we explain and illustrate how the `pim` package can be employed to fit a Probabilistic Index Model (PIM). PIMs are introduced and discussed in detail in Thas

et al. (2012) and De Neve (2013). We further illustrate the connection between PIMs and several rank-tests, as discussed in De Neve and Thas (2015). The main focus of this vignette is to illustrate the usage of the `pim` package rather than explaining PIMs in detail.

Let $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ be identically and independently distributed random vectors, where $\boldsymbol{X}$ ($\boldsymbol{X}'$) denotes the vector of covariates associated with the univariate outcome Y (Y'). A PIM is defined as:

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = g^{-1}[(\boldsymbol{X}' - \boldsymbol{X})^T \boldsymbol{\beta}], \tag{1}$$

with $\mathrm{P}\left(Y \preccurlyeq Y'\right) := \mathrm{P}\left(Y < Y'\right) + 0.5\mathrm{P}\left(Y = Y'\right)$. Here $\boldsymbol{\beta}$ denotes the parameter of interest and $g(\cdot)$ is a link function, e.g. the logit, probit or identity. Model (1) can be considered as a *standard* PIM: the right-hand side is fairly simple and we do not impose restrictions on the values of $\boldsymbol{X}$ and $\boldsymbol{X}'$. The theory developed in Thas et al. (2012) allows constructing more flexible PIMs. These PIMs can also be fitted with the `pim` package. However, for didactical purposes, we postpone this discussion to Section 3.

For notational convenience, we will sometimes drop the conditioning statement within the probability operator so that equation (1) can be simplified to:

$$\mathrm{P}\left(Y \preccurlyeq Y'\right) = g^{-1}[(\boldsymbol{X}' - \boldsymbol{X})^T \boldsymbol{\beta}].$$

2 Standard PIM

The Childhood Respiratory Disease Study (CRDS) is a longitudinal study following the pulmonary function in children. We only consider the part of this study provided by Rosner (1999). The outcome is the forced expiratory volume (FEV), which is an index of pulmonary function measured as the volume of air expelled after one second of constant effort. Along with FEV (litres), the AGE (years), HEIGHT (inches), SEX, and SMOKING status (1 if the child smokes, 0 if the child does not smoke) are provided for 654 children of ages $3-19$. See Rosner (1999, p. 41) for more information. The primary focus is on the analysis of the effect of smoking status on the pulmonary function. The data are provided with the `pim` package.

```
> library("pim")
> data(FEVData)
> head(FEVData)
```

```
  Age   FEV Height Sex Smoke
1   9 1.708   57.0   0     0
```

```
2   8 1.724   67.5   0     0
3   7 1.720   54.5   0     0
4   9 1.558   53.0   1     0
5   9 1.895   57.0   1     0
6   8 2.336   61.0   0     0
```

To model the effect of AGE and SMOKE on FEV, we consider a PIM with main effects and logit link:

$$\text{logit}\left[\mathrm{P}\left(FEV \preccurlyeq FEV'\right)\right] \;=\; \beta_1(AGE' - AGE) + \beta_2(SMOKE' - SMOKE), \quad (2)$$

where $\text{logit}(x) = \log[x/(1-x)]$. The function `pim()` can be used to fit this model and the syntax is similar to `glm()`.

```
> pim1 <- pim(formula = FEV ~ Age + Smoke, data = FEVData)
```

By default the logit-link is considered and the `pim()` function automatically translates the formula statement `FEV ~ Age + Smoke` to the formula statement of the PIM (2). More generally, a formula statement of the form $Y \sim \boldsymbol{X}$ will be automatically converted to a formula statement of the form $\mathrm{P}\left(Y \preccurlyeq Y'\right) \sim \boldsymbol{X}' - \boldsymbol{X}$.

The estimated coefficients can be extracted via `coef()`.

```
> coef(pim1)
```

```
       Age      Smoke
 0.5550350 -0.4575366
```

Consequently, $\hat{\beta}_1 = 0.56$ and $\hat{\beta}_2 = -0.46$. Inference on these parameters is obtained via the `summary()` function

```
> summary(pim1)
```

```
pim.summary of following model :
 FEV ~ Age + Smoke
Type:  difference
Link:  logit


      Estimate Std. Error z value Pr(>|z|)
```

```
Age    0.55504    0.02808  19.765   <2e-16 ***
Smoke -0.45754    0.24702  -1.852    0.064 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null hypothesis: b = 0
```

The p-values correspond to the hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$.

The estimated variance-covariance matrix of $(\hat{\beta}_1, \hat{\beta}_2)$ can be obtained with the `vcov()` function. For more functions, see `help("pim-class")`.

Thas et al. (2012) argue that the following PIM with an interaction is more appropriate:

$$\begin{aligned} \text{logit}\left[\mathrm{P}\left(FEV \preccurlyeq FEV'\right)\right] &= \beta_1(AGE' - AGE) + \beta_2(SMOKE' - SMOKE) \\ &\quad +\beta_3(AGE' * SMOKE' - AGE * SMOKE). \end{aligned} \tag{3}$$

This model can be fitted via:

```
> pim2 <- pim(FEV ~ Age*Smoke, data = FEVData)
> summary(pim2)

pim.summary of following model :
 FEV ~ Age * Smoke
Type:  difference
Link:  logit


           Estimate Std. Error z value Pr(>|z|)
Age         0.60760    0.03012  20.170  < 2e-16 ***
Smoke       5.30689    1.04423   5.082 3.73e-07 ***
Age:Smoke  -0.45539    0.07854  -5.798 6.71e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null hypothesis: b = 0
```

We end this section with an illustration of the interpretation of the effect of AGE for model (3). For 2 randomly selected children with the same smoking status (i.e. $SMOKE = SMOKE'$) and a year difference in age ($AGE = x$ and $AGE' = x + 1$), the probability that the eldest has a higher FEV is estimated by:

$$\text{expit}(0.61 - 0.46SMOKE), \quad \text{expit}(x) = \exp(x)/[1 + \exp(x)].$$

For non-smokers ($SMOKE = 0$) this probability is expit$(0.61) = 0.65$, while for smokers ($SMOKE = 1$) this becomes expit$(0.61 - 0.46) = 0.54$.

3 More complicated examples

In its most general form, a PIM is defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}_n, \tag{4}$$

where $\mathcal{X}_n$ denotes the set of pairs $(\boldsymbol{X}, \boldsymbol{X}')$ for which the model is defined. We refer to Thas et al. (2012) for more details. Model (1) is a special case where $m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}[(\boldsymbol{X}' - \boldsymbol{X})^T\boldsymbol{\beta}]$ and $\mathcal{X}_n$ does not impose any restrictions on the couples $\boldsymbol{X}$ and $\boldsymbol{X}'$. In Section 3.1 we illustrate how choices of $m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta})$ different form $g^{-1}[(\boldsymbol{X}' - \boldsymbol{X})^T\boldsymbol{\beta}]$ can be implemented and in Section 3.2 we illustrate how restrictions imposed by $\mathcal{X}_n$ can be included.

3.1 Customised formulas

To illustrate how PIMs can be fitted with customised formulas, we consider the food expenditure data set. In this study the food expenditure (FE, in Belgian francs) and the annual household income (HI, in Belgian francs) for 235 Belgian working-class households are recorded. Ernst Engel provided these data to support his hypothesis that the proportion spent on food falls with increasing income, even if actual expenditure on food rises. The data are also used in Koenker (2005) to illustrate quantile regression and are available in the `pim` and `quantreg` packages (Koenker, 2011).

```
> data(EngelData)
> head(EngelData)
```

```
    income  foodexp
1 420.1577 255.8394
2 541.4117 310.9587
3 901.1575 485.6800
4 639.0802 402.9974
5 750.8756 495.5608
6 945.7989 633.7978
```

Figure 1 indicates that the variability in food expenditure increases with increasing household income. To account for this heteroscedasticity, Thas et al. (2012) proposed

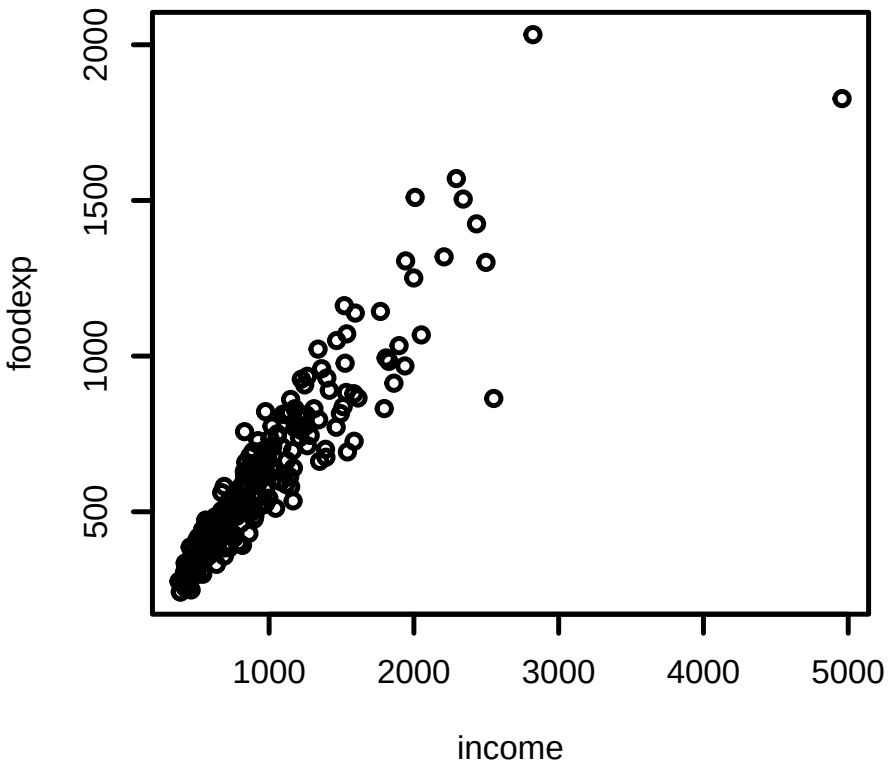

Figure 1: Food expenditure as a function of annual household income.

the following PIM:

$$\text{logit}\left[\mathrm{P}\left(FE \preccurlyeq FE'\right)\right] = \beta \frac{HI' - HI}{\sqrt{HI' + HI}}. \tag{5}$$

We refer to Thas et al. (2012) for the motivation of this model.

Because the right hand side of model (5) is not of the form $g^{-1}[(\boldsymbol{X}' - \boldsymbol{X})^T \boldsymbol{\beta}]$, we need to specify this explicitly in the formula statement upon using functions `L()` and `R()` to indicate HI and HI' respectively. Here `L()` stands for the covariate associated with the outcome at the left-hand side of the $\preccurlyeq$-sign in $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right)$, so $\boldsymbol{X}$. On the other hand, `R()` stands for the covariate associated with the outcome at the right-hand side of the $\preccurlyeq$-sign in $\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right)$, so $\boldsymbol{X}'$.

To improve readability, we shorten the names of the variables in the dataset.

```
> names(EngelData) <- c("HI", "FE")
> form <- FE ~ I( (R(HI) - L(HI))/sqrt(R(HI) + L(HI)) )
> pim3 <- pim(formula = form, data = EngelData)
> coef(pim3)

I((R(HI) - L(HI))/sqrt(R(HI) + L(HI)))
                             0.3897054
```

Similar as in `glm()` the `I()` function must be used in the formula statement to include mathematical operations.

It follows that $\hat{\beta} = 0.39$ for model (5). We briefly illustrate the interpretation. If the household income is 500 Belgian francs, the probability of larger food expenditure with a household income of 600 Belgian francs is estimated as:

$$\hat{\mathrm{P}}\left(FE \preccurlyeq FE' \mid HI = 500, HI' = 600\right) = \text{expit}\left[0.39 \frac{100}{\sqrt{500 + 600}}\right] = 0.76.$$

If we compare household incomes of 2000 and 2100 (for which the difference is also 100 Belgian francs), this effect decreases to:

$$\hat{\mathrm{P}}\left(FE \preccurlyeq FE' \mid HI = 2000, HI' = 2100\right) = \text{expit}\left[0.39 \frac{100}{\sqrt{2000 + 2100}}\right] = 0.65.$$

3.2 Restricted comparisons of the regressors

To illustrate how the $\mathcal{X}_n$ option of model (4) can be implemented, we reconsider the Childhood Respiratory Disease Study (CRDS) of Section 2. Suppose, for the sake of illustration, that one is only interested in the probability:

$$\mathrm{P}\left(FEV \preccurlyeq FEV' \mid SMOKE = 0, SMOKE' = 1, AGE = AGE'\right),$$

i.e. one wants to quantify the association between the smoking status and the pulmonary function while keeping the age fixed. Consider the PIM

$$\mathrm{P}\left(FEV \preccurlyeq FEV'\right) = \mathrm{expit}[\gamma_1 + \gamma_2(AGE' - AGE)], \tag{6}$$

where $\mathcal{X}_n$ denotes the set of pairs of children for which the first is a non-smoker, and the second is a smoker, i.e.

$$\mathcal{X}_n = \Big\{(\{SMOKE, AGE\}, \{SMOKE', AGE'\}) \mid SMOKE = 0, SMOKE' = 1\Big\}. \tag{7}$$

Note that PIM (6) is a submodel of PIM (2), but is computationally less demanding since less children have to be compared. We refer to Thas et al. (2012) for more details on $\mathcal{X}_n$ and the estimation of PIMs.

We start by construction $\mathcal{X}_n$ given by (2).

```
> id.nonsmokers <- which(FEVData$Smoke == 0)
> id.smokers <- which(FEVData$Smoke == 1)
> compare <- expand.grid(id.nonsmokers, id.smokers)
```

Next with fit the PIM (6) and give in $\mathcal{X}_n$ via the option `compare`:

```
> pim4 <- pim(formula = FEV ~ +1 + Age, data = FEVData, compare = compare)
> summary(pim4)


pim.summary of following model :
 FEV ~ +1 + Age
Type:  difference
Link:  logit


            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.05040    0.19900   0.253      0.8
Age          0.33513    0.03493   9.595   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null hypothesis: b = 0
```

Note that we explicitly have to specify the intercept. It follows that $\hat{\gamma}_1 = 0.05$ and $\hat{\gamma}_2 = 0.34$.

4 Relationship to rank tests

In this section we illustrate how several rank tests can be implemented and extended through the `pim` package. The content of this section is worked out in the appendix of De Neve and Thas (2015) using a previous, but no longer compatible, version of the package. In De Neve and Thas (2015) it is explained how the PIM can be related to well known rank tests in factorial designs and how it can be used to construct new rank tests. We start by introducing the notation used in De Neve and Thas (2015). For the factorial design with a single factor and a blocking factor we write $\boldsymbol{X} = (X, B)$, where X is a factor referring to groups or treatments of interest, and B is a blocking factor. Without loss of generality we say that X takes values $1, \ldots, K$, and B takes values $1, \ldots, L$. The number of replicates for $X = k$ and $B = l$ is denoted by n_{kl} and the total sample size is given by $N = \sum_{k=1}^{K} \sum_{l=1}^{L} n_{kl}$. Let F_{kl} denote the distribution function of Y given $X = k$ and $B = l$. In the absence of blocks, set $B = 1$ and we use the simplified notation n_k for the number of replicates for $X = k$ and F_k for the distribution function of Y given $X = k$.

Sometimes it will be convenient to work with the classical ANOVA notation. Throughout the vignette it will be clear from the context which notation is used. In particular, for $\boldsymbol{X} = (X, B)$, Y_{kl} denotes a random response variable in treatment group $k = 1, \ldots, K$ and block $l = 1, \ldots, L$. The index l becomes obsolete in the absence of blocks. We use $Y_{\cdot l}$ to denote the random response variable whose distribution is marginalized over the treatment groups, but still conditional on block l. To distinguish between the notation and model as in (4) and the ANOVA form, we refer to the former as the *regression model*, whereas models with the ANOVA notation will be referred to as the *ANOVA model*. Just like with classical linear regression models, the estimation of the parameters requires that ANOVA models are translated into regression models with dummy regressors for the coding of the factors.

4.1 Connection with the Kruskal–Wallis rank test

As a first model we define the *marginal PIM* for the K-sample layout in the absence of blocks. It is marginal in the sense that we only condition on one treatment within the PI, i.e. $\mathrm{P}(Y_i \preccurlyeq Y_j \mid X_j)$. This PI refers to the distribution of the response of observation j conditional on its regressor, i.e. $Y_j \mid X_j$, and the marginal response distribution of an observation i, i.e. Y_i. In terms of the ANOVA notation and if $X_j = k$, this becomes $\mathrm{P}(Y_{\cdot} \preccurlyeq Y_k)$, with Y_k a random response with distribution F_k and $Y_{\cdot}$ a random response with distribution $F_{\cdot} = \sum_{k=1}^{K} \lambda_k F_k$ with $\lambda_k = \lim_{N\to\infty} n_k/N$ where we assume $\lambda_k > 0$. Consider the marginal PIM in ANOVA form,

$$\mathrm{P}(Y_{\cdot} \preccurlyeq Y_k) = \alpha_k. \tag{8}$$

The interpretation of α_k is immediate: it is the probability that a random observation of group k exceeds a random observation of the marginal distribution. The corresponding PIM regression model is obtained upon defining

$$\boldsymbol{Z}_{ij}^T = \Big(\mathrm{I}\left(X_j = 1\right), \ldots, \mathrm{I}\left(X_j = K\right)\Big), \tag{9}$$

for all pairs of regressors (X_i, X_j). Let $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_K)$. Model (8) now becomes

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_j\right) = \boldsymbol{Z}_{ij}^T \boldsymbol{\alpha}, \tag{10}$$

with $\mathcal{X}_n = \{(X_i, X_j) \mid i, j = 1, \ldots, N\}$, i.e. we consider all N^2 pairs of observations.

We illustrate how model (10) can fitted to a subset of the chick weight dataset as described in Crowder and Hand (1990). Chicks are randomly allocated to one of four diets: a normal diet (referred to as diet 1) or one of three specific diets with respectively 10%, 20% or 40% protein replacement (referred to as diets 2, 3 or 4, respectively). The weights (in gram) of the chicks are measured on alternate days for the first three weeks after birth, but we only look at the weight measured at day 6 together with the weight at baseline.

```
> data(ChickWeight)
> Data <- subset(ChickWeight, Time == 6)[,-2]
> Data$baseline <- subset(ChickWeight, Time == 0)$weight[
+ is.element(subset(ChickWeight, Time == 0)$Chick, Data$Chick)]
> head(Data)
```

```
   weight Chick Diet baseline
4      64     1    1       42
16     72     2    1       40
28     67     3    1       43
40     67     4    1       42
52     60     5    1       41
64     74     6    1       41
```

Model (10) is fitted via:

```
> pim.score <- pim(formula = weight ~ R(Diet) - 1, data = Data,
+                  compare = "all",
+                  link = "identity",
+                  vcov.estim = score.vcov)
> summary(pim.score)
```

```
pim.summary of following model :
 weight ~ R(Diet) - 1
Type:  difference
Link:  identity


         Estimate Std. Error z value Pr(>|z|)
R(Diet)1  0.24507    0.05344   4.586 4.52e-06 ***
R(Diet)2  0.52396    0.08398   6.239 4.41e-10 ***
R(Diet)3  0.63125    0.08398   7.517 5.63e-14 ***
R(Diet)4  0.82917    0.08398   9.873  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null hypothesis: b = 0
```

The option `compare = "all"` indicates that all N^2 comparisons should be considered as defined by $\mathcal{X}_n$ in (10) and `vcov.estim = score.vcov` indicates that score variance-covariance matrix should be computed (i.e. the variance-covariance under the null-hypothesis $H_0 : F_1 = \ldots = F_K$). The option `link = "identity"` indicates that we fit a PIM with identity link function.

The parameters in model (8) are estimated by $\hat{\alpha}_1 = 0.25$, $\hat{\alpha}_2 = 0.52$, $\hat{\alpha}_3 = 0.63$ and $\hat{\alpha}_4 = 0.83$. Note that the p-values are associated with $H_0 : \alpha_i = 0$ which are not relevant, since they correspond to $H_0 : \mathrm{P}\left(Y_{\cdot} \preccurlyeq Y_i\right) = 0$. More relevant hypotheses are $H_0 : \alpha_i = 0.5$ and the corresponding p-values are obtained via:

```
> z.score <- (coef(pim.score) - 0.5)/sqrt(diag(vcov(pim.score)))
> 1 - pchisq(z.score^2, 1)


    R(Diet)1     R(Diet)2     R(Diet)3     R(Diet)4
1.834963e-06 7.754297e-01 1.180908e-01 8.873374e-05
```

The connection with the Kruskal–Wallis rank test is established as follows (we refer to De Neve and Thas (2015) for details):

```
> library(MASS)
> t(coef(pim.score) - 0.5)%*%ginv(vcov(pim.score))%*%c(coef(pim.score) - 0.5)


         [,1]
[1,] 28.17106
```

```
> kruskal.test(weight ~ Diet, data = Data)$stat


Kruskal-Wallis chi-squared
                  28.25611
```

The differences in both test statistics is due to ties. A Wald-type Kruskal–Wallis test statistic (using a sandwich estimator for the variance-covariance matrix of $\hat{\boldsymbol{\alpha}}$) can be obtained with the option `vcov.estim = sandwich.vcov`:

```
> pim.wald <- pim(formula = weight ~ R(Diet) - 1, data = Data,
+                   compare = "all",
+                   link = "identity",
+                   vcov.estim = sandwich.vcov)
> t(coef(pim.wald) - 0.5)%*%ginv(vcov(pim.wald))%*%c(coef(pim.wald) - 0.5)


         [,1]
[1,] 156.7046
```

4.2 Connection with the Jonckheere–Terpstra rank test

The following PIM establishes a connection with the Jonckheere–Terpstra rank test.

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_i, X_j\right) = 0.5 + \alpha[I(X_i < X_j) - I(X_i > X_j)],$$

with indicator function $I(A) = 1$ if A is true and zero otherwise. This model can be fitted employing the `R()` and `L()` arguments in the formula statement. We first order the diets.

```
> Data$Diet <- factor(Data$Diet, ordered = TRUE)
> JT.formula <- weight ~ I((L(Diet) < R(Diet)) - (L(Diet) > R(Diet))) + 1
> pim.JT <- pim(formula = JT.formula, data = Data,
+               link = "identity", vcov.estim = score.vcov,
+               compare = "all")
> summary(pim.JT)


pim.summary of following model :
 weight ~ I((L(Diet) < R(Diet)) - (L(Diet) > R(Diet))) + 1
Type:  difference
Link:  identity
```

```
                                            Estimate Std. Error   z value
(Intercept)                                5.000e-01  1.317e-09 3.796e+08
I((L(Diet) < R(Diet)) - (L(Diet) > R(Diet))) 3.678e-01  6.365e-02 5.779e+00
                                           Pr(>|z|)
(Intercept)                                 < 2e-16 ***
I((L(Diet) < R(Diet)) - (L(Diet) > R(Diet))) 7.54e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null hypothesis: b = 0
```

It follows that $\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid X_i < X_j\right)$ is estimated by $0.5 + \hat{\alpha} = 0.87$.

4.3 Connection with the Friedman rank test

The marginal PIM can be extended to block designs. In ANOVA notation this becomes

$$\mathrm{P}\left(Y_l \preccurlyeq Y_{kl}\right) = \alpha_k, \tag{11}$$

where $k = 1, \ldots, K$ refers to the treatment group and $l = 1, \ldots, L$ to the block. The interpretation of α_k is immediate: it is the probability that a random observation of group k exceeds a random observation of the marginal distribution *within the same block*. Let $\boldsymbol{Z}_{ij}$ as in (9) and $\boldsymbol{\alpha}$ as before. Model (11) in regression notation becomes

$$\mathrm{P}\left(Y_i \preccurlyeq Y_j \mid B_i, X_j, B_j\right) = \boldsymbol{Z}_{ij}^T \boldsymbol{\alpha}, \tag{12}$$

which is now only defined for $(\boldsymbol{X}_i, \boldsymbol{X}_j) \in \mathcal{X}_n = \{(\boldsymbol{X}_i, \boldsymbol{X}_j) \mid B_i = B_j, i, j = 1, \ldots, N\}$, i.e. we restrict the PI to comparisons within blocks. We refer to De Neve and Thas (2015) for more details.

To illustrate the relationship with the Friedman rank test, we consider the warpbreaks data where we consider tension as a block. This data set gives the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn, we refer to the help page of `warpbreaks` for more information. The outcome denotes the number of breaks, while the factor of interest the type of wool. The levels of tension are considered as blocks.

```
> # modify data for the sake of illustration
> wb <- aggregate(warpbreaks$breaks,
+                 by = list(w = warpbreaks$wool,
+                           t = warpbreaks$tension),
```

```
+                  FUN = mean)
> colnames(wb) = c("wool", "tension", "breaks")
> # all possible comparisons
> comp <- expand.grid(1:nrow(wb), 1:nrow(wb))
> # restrict comparisons within block
> compare <- comp[wb$tension[comp[,1]] == wb$tension[comp[,2]],]
> pim.F <- pim(breaks ~ wool, data = wb, compare = compare,
+ link = "identity",  vcov.estim = score.vcov)
> summary(pim.F)


pim.summary of following model :
 breaks ~ wool
Type:  difference
Link:  identity


      Estimate Std. Error z value Pr(>|z|)
woolB  -0.1667     0.2887  -0.577    0.564

Null hypothesis: b = 0


> friedman.test(breaks ~ wool | tension, data = wb)


        Friedman rank sum test

data:  breaks and wool and tension
Friedman chi-squared = 0.33333, df = 1, p-value = 0.5637
```

5 Remarks

Note that for a sample size of n a total $n(n-1)/2$ comparisons are considered. Consequently for large sample sizes the function goes quite slow. Bugs/comments/suggestions are welcome at Jan.DeNeve@UGent.be or Joris.Meys@UGent.be.

References

M. Crowder and D. Hand. *Analysis of Repeated Measures*. Chapman and Hal, 1990.

Jan De Neve. *Probabilistic index models*. PhD thesis, Ghent University `http://hdl.handle.net/1854/LU-3208287`, 2013.

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

R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

Roger Koenker. *quantreg: Quantile Regression*, 2011. URL `http://CRAN.R-project.org/package=quantreg`. R package version 4.54.

Bernard Rosner. *Fundamentals of Biostatistics*. Pacific Grove: Duxbury, 1999.

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