# Probabilistic Index Models 

Jan De Neve<br>Department of Data Analysis<br>Ghent University

M3 Storrs, Conneticut, USA<br>May 23, 2017

Jan.DeNeve@UGent.be

## Introduction

Introduction to Probabilistic Index Models (PIMs)

- Class of (semiparametric) regression models.
- Different than Generalized Linear Models (GLMs).
- Connection with rank-tests.
- Connection with Cox Proportional Hazards models.

Content largely based on 2 publications

- Thas, O., De Neve, J., Clement, L. and Ottoy, J.P. (2012) Probabilistic Index Models (with Discussion). JRSS-B, 74, 623-671.
- De Neve, J. and Thas, O. (2015) A Regression Framework for Rank Tests Based on the Probabilistic Index Model. JASA, 110, 1276-1283.

PIMs can be used for a variety of applications.
Current status: focus mainly on applications in biostatistics.

- Time-to-event data (survival analysis)
- Gene expression studies.


## See e.g.

De Neve, J., Thas, O., Ottoy, J.P. and Clement, L. (2013) An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data. SAGMB, 12, 333-346.

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, 2494-2495.

Goal of this talk:

Illustrate that PIMs might be useful for analyzing behavioral data

Question: What is a Probabilistic Index Model?

Question: What is a Probabilistic Index Model?
Answer: A regression model for the Probabilistic Index (PI).

Question: What is a Probabilistic Index Model?
Answer: A regression model for the Probabilistic Index (PI).
Question: What is the Probabilistic Index?

Question: What is a Probabilistic Index Model?
Answer: A regression model for the Probabilistic Index (PI).
Question: What is the Probabilistic Index?
Answer: The probability $\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)$ with $\left(Y_{i}, \boldsymbol{X}_{i}^{T}\right)$ and $\left(Y_{j}, \boldsymbol{X}_{j}\right)$ i.i.d.

## The Probabilistic Index

We will use the BtheB-study ( R package HSAUR) for motivation and illustration.

## Beat the Blues Study (BtheB):

- Clinical trial of an interactive multimedia program called Beat the Blues.
- BtheB: designed to deliver cognitive behavioural therapy to depressed patients via a computer terminal.
- Patients with depression recruited in primary care.

We will use the BtheB-study ( R package HSAUR) for motivation and illustration.

## Beat the Blues Study (BtheB):

- Clinical trial of an interactive multimedia program called Beat the Blues.
- BtheB: designed to deliver cognitive behavioural therapy to depressed patients via a computer terminal.
- Patients with depression recruited in primary care.
- Randomised to BtheB program or to 'treatment as usual' (TAU), i.e. face-to-face counselling.
- Depression is quantified via Beck Depression Inventory II (21 questions, range 0-63)
- 100 subjects in dataset (original study: 167 subjects)
- Longitudinal study, but we only consider a cross-sectional part.

Everitt and Hothorn (2015). HSAUR: A Handbook of Statistical Analyses Using R (1st Edition)
J. Proudfoot, D. Goldberg and A. Mann (2003). Computerised, interactive, multimedia CBT reduced anxiety and depression in general practice: A RCT. Psychological Medicine, 33, 217227.

## Beat the Blues Study

- Beck Depression Inventory II after 3 months (higher score $=$ more depressed).
- Beck Depression II is also measured at baseline.
- Treatment: BtheB versus TAU, randomized.
- Drugs: did the patient take anti-depressant drugs? - not randomized.
- Complete case analysis: 37 (BtheB) and 36 (TAU).


## Beat the Blues Study

- Beck Depression Inventory II after 3 months (higher score = more depressed).
- Beck Depression II is also measured at baseline.
- Treatment: BtheB versus TAU, randomized.
- Drugs: did the patient take anti-depressant drugs? - not randomized.
- Complete case analysis: 37 (BtheB) and 36 (TAU).


## Question 1

Is there a difference between the treatments in terms of depression?

## Beat the Blues Study

- Beck Depression Inventory II after 3 months (higher score $=$ more depressed).
- Beck Depression II is also measured at baseline.
- Treatment: BtheB versus TAU, randomized.
- Drugs: did the patient take anti-depressant drugs? - not randomized.
- Complete case analysis: 37 (BtheB) and 36 (TAU).


## Question 1

Is there a difference between the treatments in terms of depression?

Question 2

Does anti-depressant drug have an effect on depression?

Both treatments seem to have a positive effect.


BtheB does a slightly better job


## Modest deviation from normality

Normal QQ TAU


Normal QQ BtheB


Is there a difference between the treatments $(X)$ in terms of depression $(Y)$ ?
$H_{0}: \mathrm{E}(Y \mid X=T A U)=\mathrm{E}(Y \mid X=B$ theB $), \quad H_{A}: \operatorname{not} H_{0}$.
Two-sample t-test p-value:

- Welch: 0.041
- Permutation: 0.042 .

95\% CI for $\mathrm{E}(Y \mid X=T A U)-\mathrm{E}(Y \mid X=$ BtheB $)$ :
[0.23, 11.05]

The effect of the outlier


Results when the outlier is removed

Two-sample t-test p-value:

- Welch: 0.0083 (with outlier: 0.041)
- Permutation: 0.0087 (with outlier: 0.042).
$95 \% \mathrm{Cl}$ for $\mathrm{E}(Y \mid X=T A U)-\mathrm{E}(Y \mid X=B$ theB $)$ :

$$
[1.8,11.7] \quad \text { (with outlier: }[0.23,11.05])
$$

Since the outlier has some effect, we might want to consider a more robust test.

We choose the Wilcoxon-Mann-Whitney (WMW) Rank test

- p-value with outlier: 0.041
- p-value without outlier: 0.022

What is the effect measure associated the WMW test?

Test statistic associated with the WMW test:

$$
\begin{aligned}
& \qquad T=\frac{U-0.5}{S E_{0}(U)} \\
& U=\frac{1}{n_{B} n_{T}} \sum_{i} \sum_{j} \mathrm{I}\left(Y_{i}^{B t h e B}<Y_{j}^{T A U}\right), \mathrm{I}(T R U E)=1, \mathrm{I}(F A L S E)=0, \\
& \text { and } S E_{0}(U) \text { the standard error of } U \text { under } H_{0}: F_{B t h e B}=F_{T A U} .
\end{aligned}
$$

Test statistic associated with the WMW test:

$$
T=\frac{U-0.5}{S E_{0}(U)}
$$

$U=\frac{1}{n_{B} n_{T}} \sum_{i} \sum_{j} \mathrm{I}\left(Y_{i}^{B t h e B}<Y_{j}^{T A U}\right), \mathrm{I}(T R U E)=1, \mathrm{I}(F A L S E)=0$,
and $S E_{0}(U)$ the standard error of $U$ under $H_{0}: F_{B t h e B}=F_{T A U}$. Since

$$
\mathrm{E}(U)=\mathrm{P}\left(Y^{B t h e B}<Y^{T A U}\right)
$$

it follows that the WMW-test is associated with

$$
H_{0}: F_{B t h e B}=F_{T A U} \quad H_{A}: \mathrm{P}\left(Y^{B t h e B}<Y^{T A U}\right) \neq 0.5 .
$$

Note: under location-shift it also tests for $H_{A}: \Delta \neq 0$ with $\Delta$ a location parameter (e.g. difference in means or medians).

The effect measure

$$
\mathrm{P}\left(Y^{B t h e B}<Y^{T A U}\right)
$$

has many names:

- Mann-Whitney functional
- The nonparametric treatment effect
- The probability of superiority.
- ...
- The probabilistic index.

It is the probability that a randomly selected patient receiving BtheB will have a better (here lower) depression score than a randomly selected patient receiving TAU.

Example: $\hat{\mathrm{P}}\left(Y^{B t h e B}<Y^{T A U}\right)=64 \%(95 \% \mathrm{Cl}:[51 \%, 75 \%])$

The WMW test and the PI have some attractive properties

$$
T=\frac{U-0.5}{S E_{0}(U)} \quad \text { and } \quad \mathrm{P}\left(Y^{B t h e B}<Y^{T A U}\right)
$$

The PI

- Applies to ordinal outcomes (discrete or continuous).
- Scale-free.
- Invariant under monotone transformations of the outcome.
- 'Easy' to understand.

The WMW test

- Robust to outliers.
- Applies to ordinal outcomes (discrete or continuous).
- Good power properties: ARE.

| $\max \left(1-x^{2}, 0\right)$ | Normal | Uniform | Logistic | $t_{3}$ | Laplace | $t_{5}$ | Exp | Cauchy |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.86 | 0.95 | 1 | 1.1 | 1.24 | 1.5 | 1.9 | 3 | $\infty$ |

## Return to the Beat the Blues Study

- Beck Depression Inventory II after 3 months (higher score = more depressed).
- Beck Depression II is also measured at baseline.
- Treatment (BtheB versus TAU, randomized).
- Drugs: did the patient take anti-depressant drugs? - not randomized.
- Complete case analysis: 37 (BtheB) and 36 (TAU).


## Question 1

Is there a difference between the treatments in terms of depression?

Question 2

Does anti-depressant drug have an effect on depression?

## Does anti-depressant drug have an effect on depression?

The drugs were not randomized.


Assessing the effect of drugs on depression
Ordinary t-test: p-value $0.51,95 \% \mathrm{Cl}:[-3.9,7.6]$
$\rightarrow$ ignores the baseline score (confounder)

Assessing the effect of drugs on depression
Ordinary t-test: p-value $0.51,95 \% \mathrm{Cl}:[-3.9,7.6]$
$\rightarrow$ ignores the baseline score (confounder)

Solution: write t-test as a regression model and included baseline score as a predictor

$$
\begin{aligned}
& \qquad \operatorname{lm} \text { (score. } 3 \mathrm{M} \sim \text { drugs }+ \text { score.0M) } \\
& \quad \rightarrow \text { p-value } 0.009,95 \% \mathrm{CI}:[-7.6,-1.1] \\
& \text { (better (lower) score for those receiving drugs) }
\end{aligned}
$$

What if we are interested in the PI:

$$
\mathrm{P}\left(Y^{\text {Drugs }}<Y^{\text {No Drugs }}\right) ?
$$

Problem: Due to the confounder, we cannot trust the WMW test.
Question: Can we embed the WMW test in a regression context?

What if we are interested in the PI:

$$
\mathrm{P}\left(Y^{\text {Drugs }}<Y^{\text {No Drugs }}\right) ?
$$

Problem: Due to the confounder, we cannot trust the WMW test.
Question: Can we embed the WMW test in a regression context?
Answers: Yes, via a Probabilistic Index Model:

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=m\left(\boldsymbol{X}_{i}, \boldsymbol{X}_{j} ; \boldsymbol{\beta}\right), \quad\left(Y_{i}, \boldsymbol{X}_{i}^{T}\right) \text { i.i.d. }
$$

- $\left(Y_{i}, \boldsymbol{X}_{i}^{T}\right) i=1, \ldots, n$ i.i.d. sample
- $\boldsymbol{X}_{i}$ covariate, $p$-dimensional, e.g. $\boldsymbol{X}_{i}^{T}=$ (drugs, score.0M)
- $m(\cdot)$ a known function
- $\boldsymbol{\beta}$ the regression coefficient.


## Probabilistic Index Models

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=m\left(\boldsymbol{X}_{i}, \boldsymbol{X}_{j} ; \boldsymbol{\beta}\right),
$$

Question: how should $m\left(\boldsymbol{X}_{i}, \boldsymbol{X}_{j} ; \boldsymbol{\beta}\right)$ look like?

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=m\left(\boldsymbol{X}_{i}, \boldsymbol{X}_{j} ; \boldsymbol{\beta}\right),
$$

Question: how should $m\left(\boldsymbol{X}_{i}, \boldsymbol{X}_{j} ; \boldsymbol{\beta}\right)$ look like?

Let's have a look at the linear regression model for inspiration

$$
\mathrm{E}\left(Y_{i} \mid \boldsymbol{X}_{i}\right)=\boldsymbol{X}_{i}^{T} \boldsymbol{\beta},
$$

which implies, exploiting $\mathrm{E}\left(Y_{i}\right)-\mathrm{E}\left(Y_{j}\right)=\mathrm{E}\left(Y_{i}-Y_{j}\right)$,

$$
\mathrm{E}\left(Y_{i}-Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\left(\boldsymbol{X}_{i}-\boldsymbol{x}_{j}\right)^{\top} \boldsymbol{\beta} .
$$

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=m\left(\boldsymbol{X}_{i}, \boldsymbol{X}_{j} ; \boldsymbol{\beta}\right),
$$

Question: how should $m\left(\boldsymbol{X}_{i}, \boldsymbol{X}_{j} ; \boldsymbol{\beta}\right)$ look like?

Let's have a look at the linear regression model for inspiration

$$
\mathrm{E}\left(Y_{i} \mid \boldsymbol{X}_{i}\right)=\boldsymbol{X}_{i}^{T} \boldsymbol{\beta},
$$

which implies, exploiting $\mathrm{E}\left(Y_{i}\right)-\mathrm{E}\left(Y_{j}\right)=\mathrm{E}\left(Y_{i}-Y_{j}\right)$,

$$
\mathrm{E}\left(Y_{i}-Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\left(\boldsymbol{x}_{i}-\boldsymbol{x}_{j}\right)^{\top} \boldsymbol{\beta}
$$

So maybe the following makes sense

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=g^{-1}\left[\left(\boldsymbol{X}_{i}-\boldsymbol{X}_{j}\right)^{T} \boldsymbol{\beta}\right]
$$

with $g(\cdot)$ a link-function (e.g. probit or logit) to ensure $\mathrm{PI} \in[0,1]$.

## PIMs: connection with other models

Connection with other models.
Model 1: the parametric normal linear model:

$$
Y_{i}=\boldsymbol{X}_{i}^{T} \boldsymbol{\alpha}+\varepsilon_{i}, \quad \varepsilon_{i} \sim \mathrm{~N}\left(0, \sigma^{2}\right)
$$

Connection with other models.
Model 1: the parametric normal linear model:

$$
Y_{i}=\boldsymbol{X}_{i}^{T} \boldsymbol{\alpha}+\varepsilon_{i}, \quad \varepsilon_{i} \sim \mathrm{~N}\left(0, \sigma^{2}\right)
$$

implies

$$
\begin{aligned}
& \mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right) \\
= & \mathrm{P}\left(\boldsymbol{X}_{i}^{T} \boldsymbol{\alpha}+\varepsilon_{i}<\boldsymbol{X}_{j}^{T} \boldsymbol{\alpha}+\varepsilon_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right) \\
= & \mathrm{P}\left(\varepsilon_{i}-\varepsilon_{j}<\left(\boldsymbol{X}_{j}-\boldsymbol{X}_{i}\right)^{T} \boldsymbol{\alpha} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right) \quad \varepsilon_{i}-\varepsilon_{j} \sim \mathrm{~N}\left(0,2 \sigma^{2}\right) \\
= & \mathrm{P}\left(Z<\left(\boldsymbol{X}_{j}-\boldsymbol{X}_{i}\right)^{T} \frac{\boldsymbol{\alpha}}{\sqrt{2 \sigma^{2}}}\right) \quad Z \sim \mathrm{~N}(0,1) \\
= & g^{-1}\left[\left(\boldsymbol{X}_{j}-\boldsymbol{X}_{i}\right)^{T} \beta\right] \quad \text { with } \quad \boldsymbol{\beta}=\frac{\boldsymbol{\alpha}}{\sqrt{2 \sigma^{2}}}, \quad g(\cdot)=\operatorname{probit}(\cdot) .
\end{aligned}
$$

Connection with other models.
Model 2: semiparametric linear transformation model (part 1)

$$
h\left(Y_{i}\right)=\boldsymbol{X}_{i}^{T} \boldsymbol{\alpha}+\varepsilon_{i}, \quad \varepsilon_{i} \sim \mathrm{~N}\left(0, \sigma^{2}\right),
$$

with $h(\cdot)$ strict monotone and unknown function.

Connection with other models.
Model 2: semiparametric linear transformation model (part 1)

$$
h\left(Y_{i}\right)=\boldsymbol{X}_{i}^{T} \boldsymbol{\alpha}+\varepsilon_{i}, \quad \varepsilon_{i} \sim \mathrm{~N}\left(0, \sigma^{2}\right)
$$

with $h(\cdot)$ strict monotone and unknown function. Since

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\mathrm{P}\left(h\left(Y_{i}\right)<h\left(Y_{j}\right) \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right),
$$

if follows that

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=g^{-1}\left[\left(\boldsymbol{X}_{j}-\boldsymbol{X}_{i}\right)^{T} \boldsymbol{\beta}\right]
$$

with $\boldsymbol{\beta}=\frac{\alpha}{\sqrt{2 \sigma^{2}}}$ and $g(\cdot)=\operatorname{probit}(\cdot)$.

Connection with other models.
Model 2: semiparametric linear transformation model (part 2)
Since the difference between two extreme value variables follows a logistic distribution, one can show that

$$
h\left(Y_{i}\right)=\boldsymbol{X}_{i}^{T} \boldsymbol{\alpha}+\varepsilon_{i}, \quad \varepsilon_{i} \sim F(e)=1-\exp [-\exp (e)],
$$

implies the PIM

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=g^{-1}\left[\left(\boldsymbol{X}_{j}-\boldsymbol{X}_{i}\right)^{T} \boldsymbol{\beta}\right]
$$

with $\boldsymbol{\beta}=\boldsymbol{\alpha}$ and $g(\cdot)=\operatorname{logit}(\cdot)$.

Note: this is related to the Cox proportional hazards model.

## PIMs: estimation theory

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=g^{-1}\left[\left(\boldsymbol{X}_{j}-\boldsymbol{X}_{i}\right)^{T} \boldsymbol{\beta}\right]
$$

How can we semiparametrically estimate $\boldsymbol{\beta}$ only assuming the PIM (no further distributional assumptions)?

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=g^{-1}\left[\left(\boldsymbol{X}_{j}-\boldsymbol{X}_{i}\right)^{\top} \boldsymbol{\beta}\right],
$$

How can we semiparametrically estimate $\boldsymbol{\beta}$ only assuming the PIM (no further distributional assumptions)?

Trick:

$$
\begin{gathered}
\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\mathrm{E}\left(\iota_{i j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right), \quad \iota_{i j}=\mathrm{I}\left(Y_{i}<Y_{j}\right) \\
\Rightarrow \mathrm{E}\left(\iota_{i j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=g^{-1}\left(\boldsymbol{X}_{i j}^{T} \boldsymbol{\beta}\right), \quad \boldsymbol{X}_{i j}=\boldsymbol{X}_{j}-\boldsymbol{X}_{i}
\end{gathered}
$$

Use glm() on transformed outcomes $I_{i j}$ and predictors $\boldsymbol{X}_{i j}$ to estimate $\boldsymbol{\beta}$ !

## Challenges in the estimation

## cross-correlation:

$$
\begin{aligned}
I_{i j}=\mathrm{I}\left(Y_{i}<Y_{j}\right) & \rightarrow \mathrm{I}\left(Y_{i}<Y_{l}\right) \\
& \rightarrow \mathrm{I}\left(Y_{j}<Y_{l}\right) \\
& \rightarrow \mathrm{I}\left(Y_{k}<Y_{i}\right) \\
& \rightarrow \mathrm{I}\left(Y_{k}<Y_{j}\right)
\end{aligned}
$$

## Consequences:

- you have to prove that $\operatorname{glm}()$ gives consistent estimators.
- provide consistent sandwich estimator for $\operatorname{Var}(\hat{\boldsymbol{\beta}})$ that takes the cross-correlation into account.
- Both are solved by writing out the influence function upon using Hajek-projections.
- Nice side result: glm() does not give the efficient estimator in theory, but in practice it is very close.


## PIMs: connection with rank tests

Two-sample design

- $Y_{i}$ : depression score at 3 months.
- $X_{i}$ : anti-depressant drugs (no $=0$, yes $=1$ ).

Consider the PIM

$$
\mathrm{P}\left(Y_{i}<Y_{j} \mid X_{i}, X_{j}\right)=\operatorname{expit}\left[\left(X_{j}-X_{i}\right) \beta\right] .
$$

Two-sample design

- $Y_{i}$ : depression score at 3 months.
- $X_{i}$ : anti-depressant drugs ( $\mathrm{no}=0$, yes $=1$ ).

Consider the PIM

$$
\begin{gathered}
\mathrm{P}\left(Y_{i}<Y_{j} \mid X_{i}, X_{j}\right)=\operatorname{expit}\left[\left(X_{j}-X_{i}\right) \beta\right] \\
\rightarrow \operatorname{expit}(\beta)=\mathrm{P}\left(Y_{i}<Y_{j} \mid X_{i}=0, X_{j}=1\right)=\mathrm{P}\left(Y^{n o}<Y^{\text {yes }}\right) \\
\rightarrow \operatorname{expit}(\hat{\beta})=\frac{1}{n_{\text {no }} n_{y e s}} \sum_{i} \sum_{j} \mathrm{I}\left(Y_{i}^{\text {no }}<Y_{j}^{\text {yes }}\right)=U
\end{gathered}
$$

- Wilcoxon-Mann-Whitney test is a special case of a PIM.
- PIM sandwich estimator for $\operatorname{Var}(\hat{\beta})$ allows for Wald-type tests and the construction of confidence intervals.
- Similar results hold for the Kruskal-Wallis, Friedman, Jonckheere-Terpstra, ... rank tests.


## Return to the BtheB study

- $Y_{i}$ : depression score at 3 months.
- $X_{i}$ : anti-depressant drugs (no $=0$, yes $=1$ ).
- $Z_{i}$ : depression score at baseline.

Consider the PIM
$\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\operatorname{expit}\left[\left(X_{j}-X_{i}\right) \beta_{X}+\left(Z_{j}-Z_{i}\right) \beta_{Z}\right], \quad \boldsymbol{X}^{T}=(X, Z)$.

- $Y_{i}$ : depression score at 3 months.
- $X_{i}$ : anti-depressant drugs (no $=0$, yes $=1$ ).
- $Z_{i}$ : depression score at baseline.

Consider the PIM
$\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\operatorname{expit}\left[\left(X_{j}-X_{i}\right) \beta_{X}+\left(Z_{j}-Z_{i}\right) \beta_{Z}\right], \quad \boldsymbol{X}^{T}=(X, Z)$. In R via library('pim')
> m <- pim(bdi.3m ~ drug + bdi.pre, data = Data)
> summary (m)
pim.summary of following model :
bdi.3m ~ drug + bdi.pre
Type: difference
Link: logit
Estimate Std. Error z value $\operatorname{Pr}(>|z|)$
drugYes -0.87679 $0.31925-2.7460 .00602$ **
bdi.pre $0.08240 \quad 0.01775 \quad 4.6413 .47 \mathrm{e}-06$ ***
$\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\operatorname{expit}\left[\left(X_{j}-X_{i}\right) \beta_{X}+\left(Z_{j}-Z_{i}\right) \beta_{Z}\right], \quad \boldsymbol{X}^{T}=(X, Z)$.
From $\operatorname{pim}(): \hat{\beta}_{X}=-0.88$ and $\hat{\beta}_{Z}=0.082$

$$
\hat{\mathrm{P}}\left(Y_{i}<Y_{j} \mid X_{i}=0, X_{j}=1, Z_{i}=Z_{j}\right)=\operatorname{expit}(-0.88)=0.29
$$

The estimated probability that a patient receiving anti-depressant drugs will have a worse score (i.e. higher) as compared to a patient not receiving anti-depressant drugs is $29 \%$ ( $95 \% \mathrm{CI}$ : $[0.18,0.44]$ ).
$\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\operatorname{expit}\left[\left(X_{j}-X_{i}\right) \beta_{X}+\left(Z_{j}-Z_{i}\right) \beta_{Z}\right], \quad \boldsymbol{X}^{T}=(X, Z)$.
From $\operatorname{pim}(): \hat{\beta}_{X}=-0.88$ and $\hat{\beta}_{Z}=0.082$

$$
\hat{\mathrm{P}}\left(Y_{i}<Y_{j} \mid X_{i}=0, X_{j}=1, Z_{i}=Z_{j}\right)=\operatorname{expit}(-0.88)=0.29
$$

The estimated probability that a patient receiving anti-depressant drugs will have a worse score (i.e. higher) as compared to a patient not receiving anti-depressant drugs is $29 \%$ ( $95 \% \mathrm{CI}$ : [0.18, 0.44]). $\rightarrow$ more likely that patients receiving anti-depressant drugs will be better off.
$\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)=\operatorname{expit}\left[\left(X_{j}-X_{i}\right) \beta_{X}+\left(Z_{j}-Z_{i}\right) \beta_{Z}\right], \quad \boldsymbol{X}^{T}=(X, Z)$.
From $\operatorname{pim}(): \hat{\beta}_{X}=-0.88$ and $\hat{\beta}_{Z}=0.082$

$$
\hat{\mathrm{P}}\left(Y_{i}<Y_{j} \mid X_{i}=0, X_{j}=1, Z_{i}=Z_{j}\right)=\operatorname{expit}(-0.88)=0.29
$$

The estimated probability that a patient receiving anti-depressant drugs will have a worse score (i.e. higher) as compared to a patient not receiving anti-depressant drugs is $29 \%$ ( $95 \% \mathrm{CI}$ : [0.18, 0.44]). $\rightarrow$ more likely that patients receiving anti-depressant drugs will be better off.
$\hat{\mathrm{P}}\left(Y_{i}<Y_{j} \mid X_{i}=X_{j}, Z_{i}=z, Z_{j}=z+10\right)=\operatorname{expit}(10 \times 0.082)=0.70$.
$\rightarrow$ more likely that patients with a higher score at baseline will have a higher scare after 3 months.

## Conclusions and

 ongoing/future research
## Conclusions:

- PIMs: regression model for the Probabilistic Index $\mathrm{P}\left(Y_{i}<Y_{j} \mid \boldsymbol{X}_{i}, \boldsymbol{X}_{j}\right)$.
- Extends the Wilcoxon-Mann-Whitney test in a similar fashion as that the linear model extends the two-sample t-test.
- Estimation theory is semiparametric.
- Can be used for a variety of applications.

Ongoing/future research:

- Extend PIMs to deal with latent variables (like SEM extends linear models).
- Study what type of PIMs make sense for discrete ordinal outcomes.
- Assessing goodness-of-fit.


## References

Thas, O., De Neve, J., Clement, L. and Ottoy, J.P. (2012) Probabilistic Index Models (with Discussion). JRSS-B, 74, 623-671.

De Neve, J. and Thas, O. (2013) Goodness-of-Fit Methods for Probabilistic Index Models (213). Comm. in Stat. - Theory and Methods, 42, 1193-1207.
De Neve, J., Thas, O., Ottoy, J.P. and Clement, L. (2013) An extension of the Wilcoxon-Mann-Whitney test for analyzing RT-qPCR data. SAGMB, 12, 333-346.
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, 2494-2495.
Vermeulen, K., Thas, O., Vansteelandt, S. (2014) Increasing the power of the Mann-Whitney test in randomized experiments through flexible covariate adjustment. Stat. Med. DOI: 10.1002/sim. 6386
De Neve, J. and Thas, O. (2015) A Regression Framework for Rank Tests Based on the Probabilistic Index Model. JASA, 110, 1276-1283.
Vermeulen, K, Amorim, G, De Neve, J., Thas, O. and Vansteelandt, S. Semiparametric estimation of probabilistic index models: efficiency and bias. Under review.
Amorim, G., Thas, O., Vermeulen, K, Vansteelandt, S., and De Neve, J. Small Sample Inference for Probabilistic Index Models. Under review.
De Neve, J., Thas, O. and Gerds, T. Unified effect measures for semiparametric linear transformation models.

Meys, J, De Neve, J. and Sabbe, N. pim: Fit Probabilistic Index Models. R package version 2.0.0.2. https://CRAN.R-project.org/package=pim

> Thank you.
> Jan.DeNeve@UGent.be

