Inference on the Covariate-Specific Overlap Coefficient (OVL)

C.T. Nakas[1,2], A.M. Franco-Pereira[3], M.C. Pardo[3], B. Reiser[4]

[1]University of Thessaly
[2]Inselspital/University of Bern
[3]Complutense University of Madrid
[4]University of Haifa

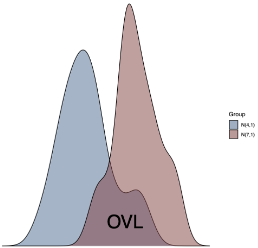
19th International Summer Conference on Probability and Statistics

The Overlap Coefficient (OVL)

- Measures similarity between two distributions:

$$\text{OVL} = \int \min(f_1(x), f_2(x))\, dx$$

- Ranges from 0 (no overlap) to 1 (identical)
- Common in diagnostic testing, bioequivalence, similarity metrics
- Not directional: unlike AUC, it captures general similarity



Overlap Coefficients - Theoretical Formulations

Matusita / Weitzman Overlap Coefficient:

$$\mathrm{OVL}(f_1, f_2) = \int \min(f_1(x), f_2(x))\, dx$$

Czekanowski / Sørensen-Dice Coefficient:

$$C(f_1, f_2) = \frac{2 \int \min(f_1(x), f_2(x))\, dx}{\int f_1(x)\, dx + \int f_2(x)\, dx}$$

Note: If f_1 and f_2 are PDFs (i.e., integrate to 1), then:

$$C(f_1, f_2) = \mathrm{OVL}(f_1, f_2)$$

Morisita Index and Comparison

Morisita Overlap Index:

$$C_M(f_1, f_2) = \frac{2\int f_1(x)f_2(x)\,dx}{\int f_1^2(x)\,dx + \int f_2^2(x)\,dx}$$

Comparison of Theoretical Overlap Indices:

Coefficient	**Symmetric**	**Range**	**Equals OVL for PDFs?**
OVL (Matusita/Weitzman)	Yes	[0, 1]	—
Czekanowski/Sørensen-Dice	Yes	[0, 1]	Yes
Morisita	Yes	[0, 1]	No

Main reference #1

Franco-Pereira AM, Nakas CT, Reiser B, Carmen Pardo M. Inference on the overlap coefficient: The binormal approach and alternatives. Stat Methods Med Res. 2021 Dec;30(12):2672-2684. doi: 10.1177/09622802211046386. Epub 2021 Oct 23. PMID: 34693817.

SageJournals

Search this journal | Enter search terms...

Advanced search

Browse by discipline | Information for

Statistical Methods in Medical Research

Impact Factor: **1.9** / 5-Year Impact Factor: **2.5**

Available access | Research article | First published online October 23, 2021

Inference on the overlap coefficient: The binormal approach and alternatives

Alba María Franco-Pereira, Christos T Nakas, [...], and María Carmen Pardo | View all authors and affiliations

Volume 30, Issue 12 | https://doi.org/10.1177/09622802211046386

Contents | PDF/EPUB | Cite article | Share options | Information, rights and permissions

Abstract

The overlap coefficient (OVL) measures the similarity between two distributions through the overlapping area of their distribution functions. Given its intuitive description and ease of visual representation by the straightforward depiction of the amount of overlap between the two corresponding histograms based on samples of measurements from each one of the two distributions, the development of accurate methods for confidence interval construction can be useful for applied researchers. The overlap coefficient has received scant attention in the literature since it lacks readily available software for its implementation, while inferential procedures that can cover the whole range of distributional scenarios for the two underlying distributions are missing. Such methods, both parametric and non-parametric are developed in this article, while R-code is provided for their implementation. Parametric approaches based on the binormal model show better performance and are appropriate for use in a wide range of distributional scenarios. Methods are assessed through a large simulation study and are illustrated using a dataset from a study on human immunodeficiency virus-related cognitive function assessment.

Estimation of OVL

Let $X_{11}, X_{12}, ..., X_{1n_1}$ and $X_{21}, X_{22}, .., X_{2n_2}$ drawn from independent $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively.

$$OVL = 1 + \Phi\left(\frac{x_1 - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{x_1 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{x_2 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{x_2 - \mu_2}{\sigma_2}\right)$$

assuming $\sigma_1^2 < \sigma_2^2$, where

$$x_1 = \frac{\left(\mu_1\sigma_2^2 - \mu_2\sigma_1^2\right) - \sigma_1\sigma_2\sqrt{\left(\mu_1 - \mu_2\right)^2 + (\sigma_1^2 - \sigma_2^2)\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)}}{(\sigma_2^2 - \sigma_1^2)}$$

$$x_2 = \frac{\left(\mu_1\sigma_2^2 - \mu_2\sigma_1^2\right) + \sigma_1\sigma_2\sqrt{\left(\mu_1 - \mu_2\right)^2 + (\sigma_1^2 - \sigma_2^2)\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)}}{(\sigma_2^2 - \sigma_1^2)}$$

where Φ is the cumulative distribution function of a $N(0, 1)$.

δ-method CIs for OVL

Denote with $\widehat{OVL}$ when plugging-in the MLEs.
Since $\widehat{\mu}_1$,$\widehat{\mu}_2$, $\widehat{\sigma}_1$, $\widehat{\sigma}_2$ are all independent, using the delta method, we obtain

$$Var\left(\widehat{OVL}\right) \approx \left(\frac{\partial OVL}{\partial \mu_1}\right)^2 Var\left(\widehat{\mu}_1\right) + \left(\frac{\partial OVL}{\partial \mu_2}\right)^2 Var\left(\widehat{\mu}_2\right) + \left(\frac{\partial OVL}{\partial \sigma_1}\right)^2 Var\left(\widehat{\sigma}_1\right) + \left(\frac{\partial OVL}{\partial \sigma_2}\right)^2 Var\left(\widehat{\sigma}_2\right),$$

where the derivatives are evaluated using the MLEs

The δ-method CIs are

$$\widehat{OVL} \pm z_{1-\alpha/2}\sqrt{Var\left(\widehat{OVL}\right)}$$

The Box-Cox transformation

In the binormal framework:

$$X_{ij_i}^{(\lambda)} = \begin{cases} \dfrac{X_{ij_i}^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \log(X_{ij_i}) & \lambda = 0 \end{cases}$$

The MLE of the common transformation parameter λ can be obtained by maximizing the profile likelihood function given by

$$l(\lambda) = -\frac{n_1}{2}\log\left(\frac{\sum_{i=1}^{n_1}\left(x_{1i}^{(\lambda)} - \frac{\sum_{i=1}^{n_1} x_{1i}^{(\lambda)}}{n_1}\right)^2}{n_1}\right) - \frac{n_2}{2}\log\left(\frac{\sum_{j=1}^{n_2}\left(x_{2j}^{(\lambda)} - \frac{\sum_{j=1}^{n_2} x_{2j}^{(\lambda)}}{n_2}\right)^2}{n_2}\right) + (\lambda - 1)\Big(\sum_{i=1}^{n_1}\log X_{1i} + \sum_{j=1}^{n_2}\log X_{2j}\Big) + k,$$

where k is constant.

δ-BC and δ-BC-λ CIs for *OVL*

There are two possible ways to proceed:

- δ-BC: act as if λ is known, and then use its estimated value to transform the observations and compute the confidence intervals.

- δ-BC-λ: Take into account the variability of the estimated λ by analyzing the full likelihood function. We plug-in $\widehat{\Sigma}^{(\lambda)}$, an estimator of the covariance matrix of the estimated parameter vector $\left(\mu_1^{(\lambda)}, \mu_2^{(\lambda)}, \sigma_1^{(\lambda)}, \sigma_2^{(\lambda)}\right)$:

$$Var_\lambda\left(\widehat{OVL}^{(BC)}\right) \approx \left(\frac{\partial \widehat{OVL}^{(BC)}}{\partial \mu_1}, \frac{\partial \widehat{OVL}^{(BC)}}{\partial \mu_2}, \frac{\partial \widehat{OVL}^{(BC)}}{\partial \sigma_1}, \frac{\partial \widehat{OVL}^{(BC)}}{\partial \sigma_2}\right) \widehat{\Sigma}^{(\lambda)}$$

$$\left(\frac{\partial \widehat{OVL}^{(BC)}}{\partial \mu_1}, \frac{\partial \widehat{OVL}^{(BC)}}{\partial \mu_2}, \frac{\partial \widehat{OVL}^{(BC)}}{\partial \sigma_1}, \frac{\partial \widehat{OVL}^{(BC)}}{\partial \sigma_2}\right)^t$$

The Logit scale

One can switch to the logit scale,

$$logit\left(\widehat{OVL}\right) \pm z_{1-\alpha/2}\sqrt{Var\left(logit\left(\widehat{OVL}\right)\right)},$$

where $logit\left(\widehat{OVL}\right) = \log\left(\widehat{OVL}/\left(1-\widehat{OVL}\right)\right)$ and
$Var\left(logit\left(\widehat{OVL}\right)\right) = Var\left(\widehat{OVL}\right)/\left(\widehat{OVL}\left(1-\widehat{OVL}\right)\right)^2$

Transforming back to the original scale, confidence intervals for OVL can be obtained that are properly bounded in $[0,1]$.

We get "L-δ", "L-δ-BC" and "L-δ-BC-λ"

Bootstrap CIs

BC-AN: Another possibility is to consider a parametric bootstrap approach to estimate the variance of the OVL estimator:

1.- Sample with replacement from X_1 and X_2

2.- Box-Cox transformation for each bootstrap sample

3.- For $i = 1, 2$, calculate $\widehat{\mu}_i$ and $\widehat{\sigma}_i$

4.- Calculate $\widehat{OVL}$

5.- Repeat 1-4 B times.

Then, based on the B bootstrapped values of $\widehat{OVL}$, $\widehat{OVL}_b^*$, derive the bootstrap estimate of the variance $Var_B\left(\widehat{OVL}\right)$

$$\widehat{OVL} \pm z_{1-\alpha/2}\sqrt{Var_B\left(\widehat{OVL}\right)}$$

- L-BC-AN: logit scale and back-transformed
- BC-PB: Bootstrap percentile CI
- BC-bias: Bias corrected bootstrap CI

Non-parametric approaches

The normality assumption may not be satisfied even after applying the Box-Cox transformation. Kernel-based approaches were considered. The density estimator for $f_{X_1}(x)$ is given by

$$\widehat{f}_{X_1}(x) = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{h} K\left(\frac{x - X_{in_1}}{h}\right)$$

where K is a kernel function and the bandwidth is given by

$$h = \left(\frac{4}{3}\right)^{1/5} s n_1^{-1/5},$$

with

$$s = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_{in_1} - \sum_{j=1}^{n_1} \frac{X_{jn_1}}{n_1}\right)^2}.$$

Define $\widehat{f}_{X_2}(x)$ analogously and proceed via bootstrap to estimate the variance. We denote this estimator $\widehat{OVL}^{(K)}$.

Basic kernel approach

K-NSR: Confidence intervals for OVL can be obtained via the following steps:

1.- Draw bootstrap samples of sizes n_1 and n_2 with replacement from X_1 and X_2, respectively. Denote these by X_{1b} and X_{2b}.

2.- Obtain the kernel based OVL estimates, denoted by $\widehat{OVL}_b^{(K)}$, for the current (b^{th}) boostrap samples X_{1b} and X_{2b}.

3.- Repeat Steps 1-2 B times and derive the boostrap-based estimate of the variance of $\widehat{OVL}^{(K)}$ by calculating $Var\left(\widehat{OVL}^{(K)}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{OVL}_b^{(K)} - \overline{\widehat{OVL}}^{(K)}\right)$ where $\overline{\widehat{OVL}}^{(K)}$ is the mean of the B terms $\widehat{OVL}_b^{(K)}$ $(b = 1, ..., B)$.

4.- Construct the two-sided $100(1-\alpha)\%$ confidence interval of OVL via $\overline{\widehat{OVL}}^{(K)} \pm z_{1-\alpha/2}\sqrt{Var\left(\widehat{OVL}^{(K)}\right)}$.

Other kernel-based variants

- K-CV: Same but considering the univariate smoothed cross-validation bandwidth selector.
- Working on the logit scale again, the different approaches are "L-K-NSR" or "L-K-CV" depending on the bandwidth.
- Finally, the bootstrap estimates $\widehat{OVL}_b^*$ can also be used to obtain bootstrap percentile-based confidence intervals ("K–NSR-PB") and ("K–CV-PB").

In all these cases we have considered two different kernel functions K: the Gaussian and the Epanechnikov kernels (and preferred the latter).

Method overview 1/3

δ	$\widehat{OVL} \pm z_{1-\alpha/2}\sqrt{Var\left(\widehat{OVL}\right)}$	Parametric approach using the delta method
δ-BC	$\widehat{OVL}^{(BC)} \pm z_{1-\alpha/2}\sqrt{Var\left(\widehat{OVL}^{(BC)}\right)}$	Parametric approach using the delta method after the Box-Cox transformation
δ-BC-λ	$\widehat{OVL}^{(BC)} \pm z_{1-\alpha/2}\sqrt{Var_{\lambda}\left(\widehat{OVL}^{(BC)}\right)}$	Parametric approach using the delta method after the Box-Cox transformation taking into account the variability of the estimated transformation parameter
L-δ	$\frac{exp\left(logit\left(\widehat{OVL}\right)\pm z_{1-\alpha/2}\sqrt{Var\left(logit\left(\widehat{OVL}\right)\right)}\right)}{1+exp\left(logit\left(\widehat{OVL}\right)\pm z_{1-\alpha/2}\sqrt{Var\left(logit\left(\widehat{OVL}\right)\right)}\right)}$	Parametric approach using the delta method after switching to a logit scale and then transforming back
L-δ-BC	$\frac{exp\left(logit\left(\widehat{OVL}^{(BC)}\right)\pm z_{1-\alpha/2}\sqrt{Var\left(logit\left(\widehat{OVL}^{(BC)}\right)\right)}\right)}{1+exp\left(logit\left(\widehat{OVL}^{(BC)}\right)\pm z_{1-\alpha/2}\sqrt{Var\left(logit\left(\widehat{OVL}^{(BC)}\right)\right)}\right)}$	Parametric approach using the delta method after the Box-Cox transformation after switching to a logit scale and then transforming back

Method overview 2/3

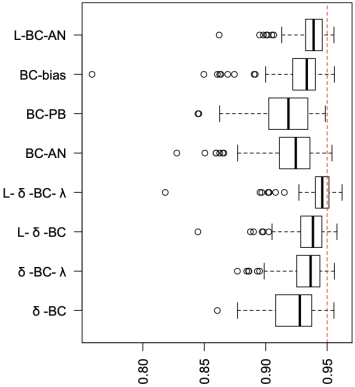
L-δ-BC-λ	$\frac{exp\left(logit\left(\widehat{OVL}^{(BC)}\right)\pm z_{1-\alpha/2}\sqrt{Var_\lambda\left(logit\left(\widehat{OVL}^{(BC)}\right)\right)}\right)}{1+exp\left(logit\left(\widehat{OVL}^{(BC)}\right)\pm z_{1-\alpha/2}\sqrt{Var_\lambda\left(logit\left(\widehat{OVL}^{(BC)}\right)\right)}\right)}$	Parametric approach using the delta method after the Box-Cox transformation in the logit scale and back-transformed considering the variability of the estimated transformation parameter
BC-AN	$\widehat{OVL}^{(BC)} \pm z_{1-\alpha/2}\sqrt{Var_B\left(\widehat{OVL}^{(BC)}\right)}$	Parametric approach using a bootstrap-based approach to estimate the variance
BC-PB	$(\widehat{OVL}^*_{(1-\alpha/2)}, \widehat{OVL}^*_{(\alpha/2)})$	Parametric approach using a bootstrap percentile approach
BC-bias	$(\widehat{OVL}^*_{(\alpha_1)}, \widehat{OVL}^*_{(\alpha_2)})$	Parametric approach using a bootstrap bias-corrected approach

Method overview 3/3

L-BC-AN	$\dfrac{exp\left(logit\left(\widehat{OVL}^{(BC)}\right) \pm z_{1-\alpha/2}\sqrt{Var_B\left(logit\left(\widehat{OVL}^{(BC)}\right)\right)}\right)}{1+exp\left(logit\left(\widehat{OVL}^{(BC)}\right) \pm z_{1-\alpha/2}\sqrt{Var_B\left(logit\left(\widehat{OVL}^{(BC)}\right)\right)}\right)}$	BC-AN procedure carried out in the logit scale and back-transformed
K-NSR K-CV	$\widehat{OVL}^{(K)} \pm z_{1-\alpha/2}\sqrt{Var\left(\widehat{OVL}^{(K)}\right)}$	Kernel approach estimating the variance via bootstrap
L-K-NSR L-K-CV	$\dfrac{exp\left(logit\left(\widehat{OVL}^{(K)}\right) \pm z_{1-\alpha/2}\sqrt{Var\left(logit\left(\widehat{OVL}^{(K)}\right)\right)}\right)}{1+exp\left(logit\left(\widehat{OVL}^{(K)}\right) \pm z_{1-\alpha/2}\sqrt{Var\left(logit\left(\widehat{OVL}^{(K)}\right)\right)}\right)}$	Kernel approach estimating the variance via bootstrap in the logit scale and back-transformed
K-NSR-PB K-CV-PB	$(\widehat{OVL}^{(K)*}_{(\alpha_1)}, \widehat{OVL}^{(K)*}_{(\alpha_2)})$	Kernel approach using a bootstrap percentile approach

Some simulations results

All parametric simulations considered...



R-package OVL.CI

OVL.CI: Inference on the Overlap Coefficient: The Binormal Approach and Alternatives

Provides functions to construct confidence intervals for the Overlap Coefficient (OVL). OVL measures the similarity between two distributions through the overlapping area of their distribution functions. Given its intuitive description and ease of visual representation by the straightforward depiction of the amount of overlap between the two corresponding histograms based on samples of measurements from each one of the two distributions, the development of accurate methods for confidence interval construction can be useful for applied researchers. Implements methods based on the work of Franco-Pereira, A.M., Nakas, C.T., Reiser, B., and Pardo, M.C. (2021) <doi:10.1177/09622802211046386>.

Version:	0.1.0
Depends:	R (≥ 2.10)
Imports:	ks
Suggests:	testthat (≥ 3.0.0)
Published:	2023-11-13
DOI:	10.32614/CRAN.package.OVL.CI
Author:	Alba M. Franco-Pereira [aut, cre, cph], Christos T. Nakas [aut], Benjamin Reiser [aut], M.Carmen Pardo [aut]
Maintainer:	Alba M. Franco-Pereira <albfranc at ucm.es>
License:	GPL-2
NeedsCompilation:	no
Language:	en-US
Materials:	NEWS
CRAN checks:	OVL.CI results

Documentation:

Reference manual: OVL.CI.pdf

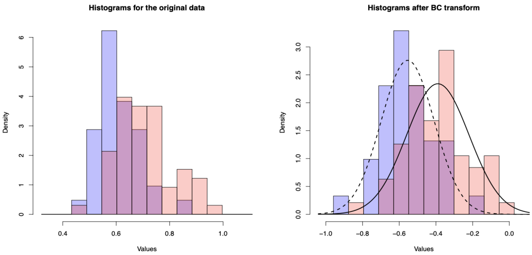
Downloads:

Package source:	OVL.CI_0.1.0.tar.gz
Windows binaries:	r-devel: OVL.CI_0.1.0.zip, r-release: OVL.CI_0.1.0.zip, r-oldrel: OVL.CI_0.1.0.zip
macOS binaries:	r-release (arm64): OVL.CI_0.1.0.tgz, r-oldrel (arm64): OVL.CI_0.1.0.tgz, r-release (x86_64): OVL.CI_0.1.0.tgz, r-oldrel (x86_64): OVL.CI_0.1.0.tgz

Linking:

Please use the canonical form `https://CRAN.R-project.org/package=OVL.CI` to link to this page.

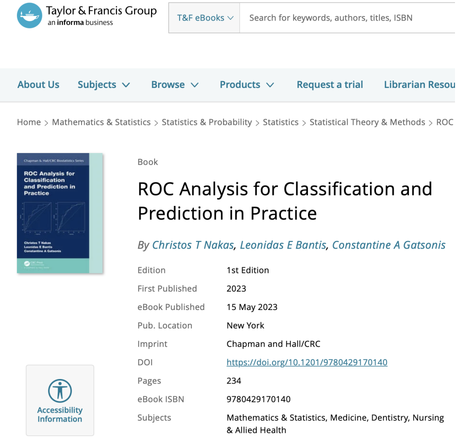
Myoinositol (MI) over creatine (Cr) measured via proton MRS was used from a published study involving 39 neurologically asymptomatic (NAS) HIV+ patients and 37 HIV– controls (NEG), as a biomarker for dementia.



Method	Point Estimate	95% CI LC	95% CI UC	Width
L-δ-BC-λ	0.5751	0.4390	0.7313	0.2923
L-BC-AN	0.5934	0.4360	0.7337	0.2977

ROC analysis?

Nakas, C.T., Bantis, L.E., & Gatsonis, C.A. (2023). ROC Analysis for Classification and Prediction in Practice (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429170140



Why Covariate Adjustment?

- OVL is often affected by patient characteristics (e.g. age) since covariates may significantly influence distributional overlap.
- Ignoring covariates can mask heterogeneity.
- Need to estimate OVL(z): overlap conditional on covariates using linear regression with a possible Box-Cox transformation.
- We develop procedures for CI calculation for OVL(z).

Main reference #2

Accepted in *Journal of Biopharmaceutical Statistics*

Confidence intervals for the covariate-specific overlap coefficient (OVL)

M. Carmen Pardo[1,2], Alba M. Franco-Pereira[1,2,*], Benjamin Reiser[3] and Christos T. Nakas[4,5]

Model Framework

- Linear regression framework: $X_i = \beta^T Z + \varepsilon$
- Conditional densities $f_1(x|z), f_2(x|z)$
- Box-Cox transformation to normalize data if needed
- OVL(z): overlap between conditional densities

Basic model components

- Let $\left\{\left(\mathrm{z}_{\overline{D}_i}, x_{\overline{D}_i}\right)\right\}_{i=1}^{n_{\overline{D}}}$ and $\{(\mathrm{z}_{D_i}, x_{D_i})\}_{i=1}^{n_D}$ be two independent random samples of test outcomes and covariates from non-diseased and diseased groups of size $n_{\overline{D}}$ and n_D, respectively.
- Where $\mathrm{z}_{\overline{D}_i} = \left(z_{\overline{D}_{i,1}}, \ldots, z_{\overline{D}_{i,p_1}}\right)^t$ and $\mathrm{z}_{D_i} = \left(z_{D_{i,1}}, \ldots, z_{D_{i,p_2}}\right)^t$ be p_1-dimensional and p_2-dimensional vectors of covariates, respectively.
- Assume that the (possibly transformed) non-diseased and diseased marker values can be described as linear functions of the explanatory variables.

$$\begin{aligned} \mathrm{X}_{\overline{D}} &= \mathrm{Z}_{\overline{D}} \boldsymbol{\beta}_{\overline{D}} + \epsilon_{\overline{D}}, \\ \mathrm{X}_D &= \mathrm{Z}_D \boldsymbol{\beta}_D + \epsilon_D, \end{aligned}$$

- where $\mathrm{X}_{\overline{D}} = \left(x_{\overline{D}_1}, \ldots, x_{\overline{D}_{n_{\overline{D}}}}\right)^t$, $\mathrm{X}_D = \left(x_{D_1}, \ldots, x_{D_{n_D}}\right)^t$ and $\boldsymbol{\beta}_{\overline{D}}$ and $\boldsymbol{\beta}_D$ are the parameters' column vectors of size p_1 and p_2, respectively.

OVL(z) estimation 1/2

The covariate adjusted OVL for a vector of covariates $\mathrm{Z}=\mathrm{z}$ of dimension p is given by

$$OVL\left(\mathrm{z}\right)=1+\Phi\left(\frac{x_1\left(\mathrm{z}\right)-\boldsymbol{\beta}_{\bar{D}}^t\mathrm{z}}{\sigma_{\bar{D}}}\right)-\Phi\left(\frac{x_1\left(\mathrm{z}\right)-\boldsymbol{\beta}_{D}^t\mathrm{z}}{\sigma_{D}}\right)-\Phi\left(\frac{x_2\left(\mathrm{z}\right)-\boldsymbol{\beta}_{\bar{D}}^t\mathrm{z}}{\sigma_{\bar{D}}}\right)+\Phi\left(\frac{x_2\left(\mathrm{z}\right)-\boldsymbol{\beta}_{D}^t\mathrm{z}}{\sigma_{D}}\right)$$

assuming that $\sigma_{\bar{D}}^2<\sigma_{D}^2$, where:

$$x_1\left(\mathrm{z}\right)=\frac{\left(\boldsymbol{\beta}_{\bar{D}}^t\mathrm{z}\sigma_D^2-\boldsymbol{\beta}_{D}^t\mathrm{z}\sigma_{\bar{D}}^2\right)-\sigma_{\bar{D}}\sigma_D\sqrt{\left(\boldsymbol{\beta}_{\bar{D}}^t\mathrm{z}-\boldsymbol{\beta}_{D}^t\mathrm{z}\right)^2+\left(\sigma_{\bar{D}}^2-\sigma_D^2\right)\log\left(\frac{\sigma_{\bar{D}}^2}{\sigma_D^2}\right)}}{\left(\sigma_D^2-\sigma_{\bar{D}}^2\right)}$$

and

$$x_2\left(\mathrm{z}\right)=\frac{\left(\boldsymbol{\beta}_{\bar{D}}^t\mathrm{z}\sigma_D^2-\boldsymbol{\beta}_{D}^t\mathrm{z}\sigma_{\bar{D}}^2\right)+\sigma_{\bar{D}}\sigma_D\sqrt{\left(\boldsymbol{\beta}_{\bar{D}}^t\mathrm{z}-\boldsymbol{\beta}_{D}^t\mathrm{z}\right)^2+\left(\sigma_{\bar{D}}^2-\sigma_D^2\right)\log\left(\frac{\sigma_{\bar{D}}^2}{\sigma_D^2}\right)}}{\left(\sigma_D^2-\sigma_{\bar{D}}^2\right)}.$$

OVL(z) estimation 2/2

$OVL\,(\mathrm{z})$ is estimated by plugging-in the relevant estimates for the unknown $\boldsymbol{\beta}_{\overline{D}}, \boldsymbol{\beta}_D, \sigma^2_{\overline{D}}, \sigma^2_D$ in the formulae given in the previous slide:

$$\widehat{\boldsymbol{\beta}}_{\overline{D}} = \left(\mathrm{Z}^t_{\overline{D}}\mathrm{Z}_{\overline{D}}\right)^{-1}\mathrm{Z}^t_{\overline{D}}\mathrm{X}_{\overline{D}}; \qquad \widehat{\boldsymbol{\beta}}_D = \left(\mathrm{Z}^t_D\mathrm{Z}_D\right)^{-1}\mathrm{Z}^t_D\mathrm{X}_D;$$

$$\widehat{\sigma}^2_{\overline{D}} = \frac{\left(\mathrm{X}^t_{\overline{D}}\mathrm{X}_{\overline{D}} - \widehat{\boldsymbol{\beta}}^t_{\overline{D}}\mathrm{Z}^t_{\overline{D}}\mathrm{X}_{\overline{D}}\right)}{n_{\overline{D}} - p}; \qquad \widehat{\sigma}^2_D = \frac{\left(\mathrm{X}^t_D\mathrm{X}_D - \widehat{\boldsymbol{\beta}}^t_D\mathrm{Z}^t_D\mathrm{X}_D\right)}{n_D - p}.$$

Confidence Intervals for OVL(z)

- Delta method: analytic variance (complex)
- Bootstrap methods:
  - PB (percentile)
  - Bias-corrected (bias)
  - Logit + AN using bootstrap variance estimate (L-AN, LD-AN)
  - Box-Cox extensions (L-BC-AN, LD-BC-AN, BC-PB, BC-bias)
- Logit transformation helps bound CI in [0,1]

Basic bootstrap CIs for OVL(z)

Method L-AN:

- Compute $\widehat{\beta}^t_{\overline{D}}$, $\widehat{\beta}^t_D$, $\widehat{\sigma}^2_{\overline{D}}$, $\widehat{\sigma}^2_D$ from the original sample to calculate $\widehat{OVL\,(\mathrm{z})}$
- Sample with replacement from $(\mathrm{X}_{\overline{D}}, Z_{\overline{D}})$ and (X_D, Z_D) and calculate $\widehat{OVL\,(\mathrm{z})}$ for the bootstrap sample.
- Repeat B times. Then, based on the B bootstrapped values $\widehat{OVL}_b(\mathrm{z})$, derive the bootstrap estimate of $Var_B\left(\widehat{OVL\,(\mathrm{z})}\right)$.
- Construct the two-sided $100(1-\alpha)\%$ confidence interval of $OVL\,(\mathrm{z})$ switching to the logit scale and then transforming back to the original scale, through

$$logit\left(\widehat{OVL\,(\mathrm{z})}\right) \pm z_{1-\alpha/2}\sqrt{Var\left(logit\left(\widehat{OVL\,(\mathrm{z})}\right)\right)}.$$

Bootstrap CI variants

Alternatives:

- LD-AN: Estimate the variance of the logit of OVL using the bootstrap estimate of logit OVL directly.
- The bootstrap estimates, $\widehat{OVL}_b(z)$ can also be used to obtain bootstrap percentile confidence intervals ('PB') as well as the bias corrected bootstrap confidence interval ('bias').

Box-Cox variants

Box-Cox transformation:

$$x_{\overline{D}}^{(\lambda)} \mid z = \begin{cases} \dfrac{(x_{\overline{D}}^{(\lambda)} \mid z) - 1}{\lambda} & \lambda \neq 0 \\[2ex] \log(x_{\overline{D}}^{(\lambda)} \mid z) & \lambda = 0 \end{cases} \qquad x_{D}^{(\lambda)} \mid z = \begin{cases} \dfrac{(x_{D}^{(\lambda)} \mid z) - 1}{\lambda} & \lambda \neq 0 \\[2ex] \log(x_{D}^{(\lambda)} \mid z) & \lambda = 0 \end{cases}$$

where the maximum likelihood estimate of the common transformation parameter λ can be obtained by maximizing the log-profile likelihood function given by

$$l(\lambda) = \sum_{i \in \{\overline{D}, D\}} \left[-\frac{n_i}{2} \left(\log 2\pi \widehat{\sigma}_i^2\right) - \frac{1}{2} \sum_{j=1}^{n_i} \left(\left(x_{i_j}^{(\lambda)} - \widehat{\beta}_i^t z_i \right)^2 / \widehat{\sigma}_i^2 \right) + (\lambda - 1) \sum_{j=1}^{n_i} \log x_{i_j}^{(\lambda)} \right]$$

Notes on the Box-Cox

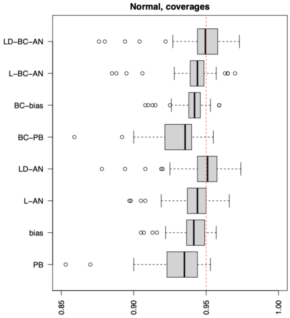
- When the Box-Cox transformation is used for the regression model, we add a step 0 in the bootstrap algorithm using the transformation first on the original sample data.
- To obtain the bootstrapped estimates and compute the variance, we need to carry out a Box-Cox transformation between steps 2 and 3. Specifically, after sampling with replacement from the original sample, we apply the BC transformation.
- Methods denoted by L-BC-AN, LD-BC-AN, BC-PB and BC-bias then arise.
- The Box-Cox transformation leads to a nonlinear relationship between the marker of interest and the covariates.

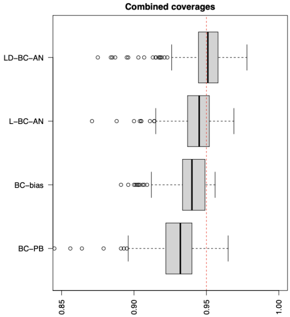
Simulation Study Design

- Scenarios: Normal, LogNormal, PowerNormal
- Sample sizes: 50, 100, 200, 500
- Metrics: Coverage probability, CI width
- Methods compared: PB, bias, L-AN, LD-AN, L-BC-AN, LD-BC-AN, BC-PB, BC-bias

Key Result: LD-BC-AN Performs Best overall

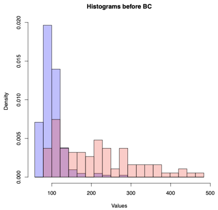
- LD-BC-AN: stable, accurate, and robust
- Box-Cox adds robustness with minimal cost under normality
- PB often undercovers; logit-based methods superior

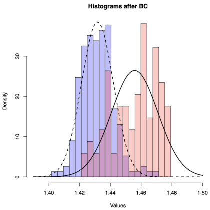


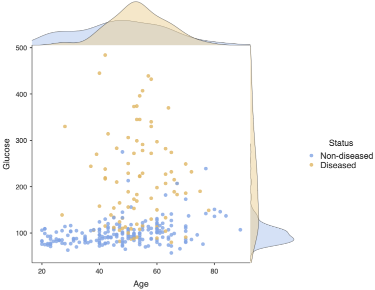


Real Data: Glucose as a Diabetes marker adjusted for Age

- 286 subjects (198 non-diseased, 88 diseased)
- Post-prandial glucose as marker for diabetes
- Age as covariate

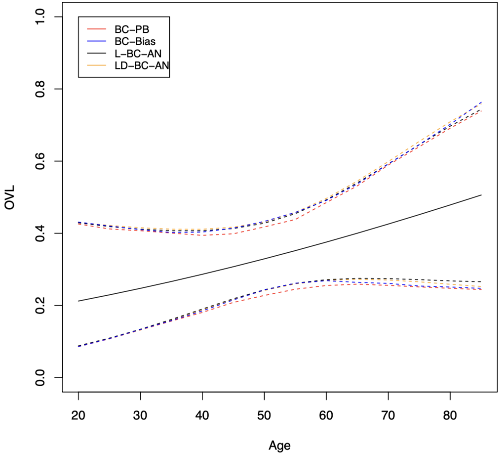






OVL(Age): Covariate-Specific Insight

- Better discrimination in younger subjects
- Age-adjusted OVL provides meaningful stratified insight



Summary

- Introduced inference for covariate-specific OVL
- Logit + bootstrap-based methods work best
- Box-Cox transformation improves flexibility
- LD-BC-AN is a recommended method
- Future steps: Incorporate into R package ('OVL.CI')

Thank You

Questions?

`cnakas@uth.gr`