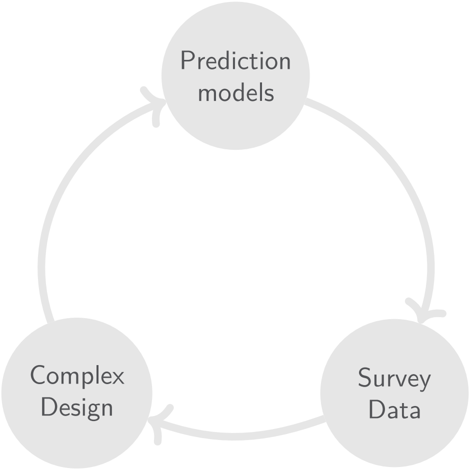
Addressing Complex Sampling Designs in the Development of Regression Models

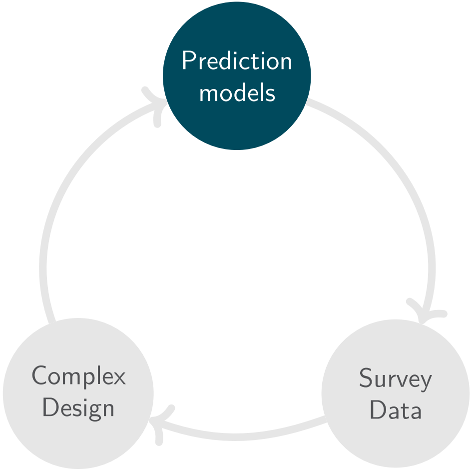
Amaia Iparragirre[1], Irantzu Barrio[1,2], Inmaculada Arostegui[1,2]

[1] University of the Basque Country (UPV/EHU)
[2] BCAM-Basque Center for Applied Mathematics

Universidad de Navarra | DATAI
INSTITUTO DE CIENCIA DE LOS
DATOS E INTELIGENCIA ARTIFICIAL

1. Introduction
2. Methodological proposals
3. Software development
4. Discussion and further research


Prediction models
Survey Data
Complex Design


Prediction
models
Complex
Design
Survey
Data

Prediction models

Based on known past behavior...

... what is most likely to happen in the future?

Purpose: To make future predictions by means of known past results.

Prediction models

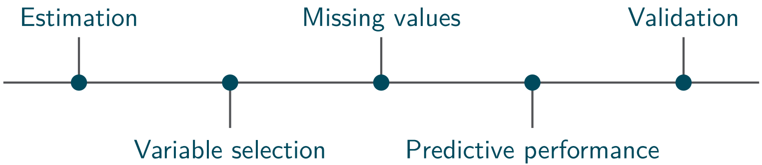
Based on known past behavior...

... what is most likely to happen in the future?

Purpose: To make future predictions by means of known past results.

Development of prediction models

Several steps should be considered in the development of prediction models in order to end up with a **valid model**:



Prediction models

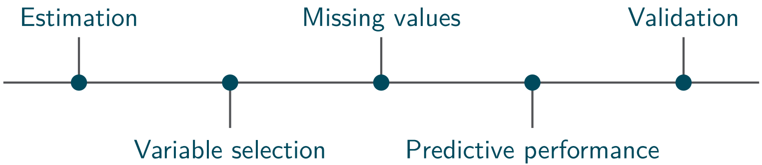
Based on known past behavior...

... what is most likely to happen in the future?

Purpose: To make future predictions by means of known past results.

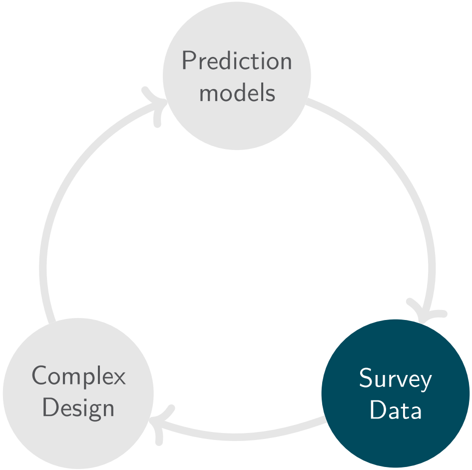
Development of prediction models

Several steps should be considered in the development of prediction models in order to end up with a **valid model**:

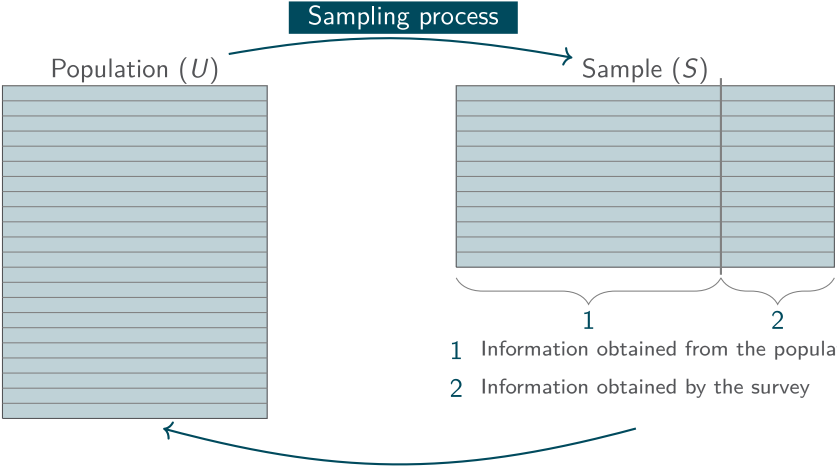


Existing techniques: data need to satisfy iid conditions

Survey data



Survey data

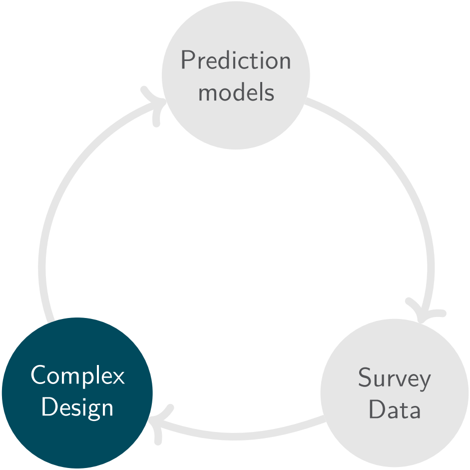


1 Information obtained from the population
2 Information obtained by the survey

The goal is to make conclusions related to the population based on the information obtained by means of the survey

Prediction models

Motivation

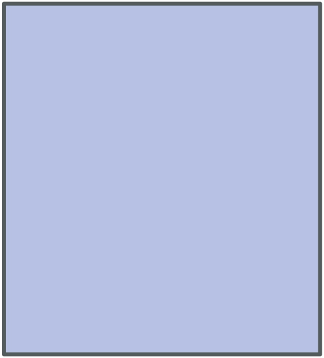


Complex sampling designs

One-stage stratified sampling

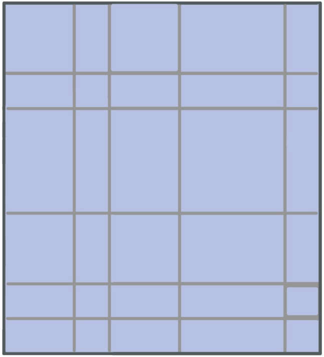
Complex sampling designs

One-stage stratified sampling



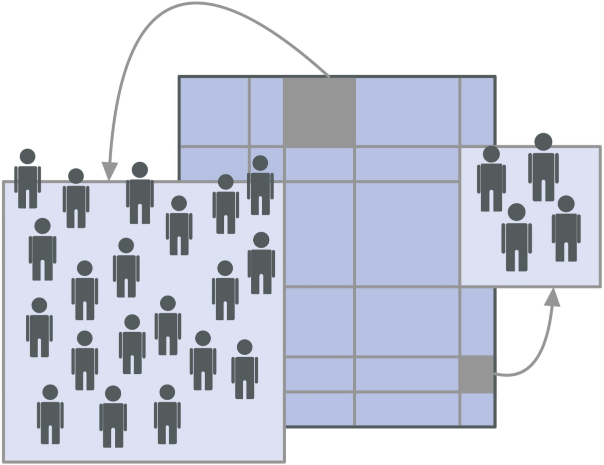
Complex sampling designs

One-stage stratified sampling



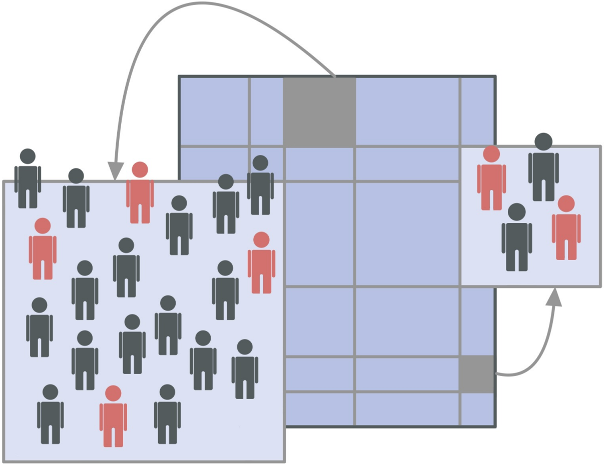
Complex sampling designs

One-stage stratified sampling



Complex sampling designs

One-stage stratified sampling



Complex sampling designs

One-stage stratified sampling

Population U (of size N):

$$U = \bigcup_{h=1}^{H} U_h \text{ , each } U_h \text{ of size } N_h, \ \forall h \in \{1, \dots, H\}.$$

Inclusion probabilities:

$$\pi_i = \frac{n_h}{N_h}, \quad \forall i \in U_h, \quad \forall h \in \{1, \dots, H\}.$$

Sampling weights

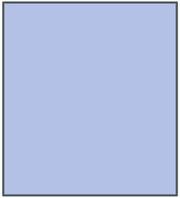
$$w_i = \frac{1}{\pi_i} = \frac{N_h}{n_h}, \quad \forall i \in S_h, \quad \forall h \in \{1, \dots, H\}, \ S = \cup_{h=1}^{H} S_h.$$

Complex sampling designs

Two-stage stratified cluster sampling

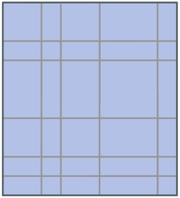
Complex sampling designs

Two-stage stratified cluster sampling



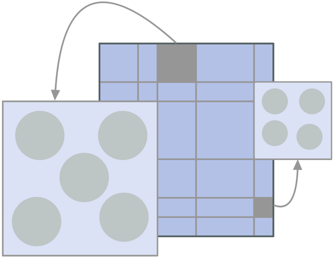
Complex sampling designs

Two-stage stratified cluster sampling



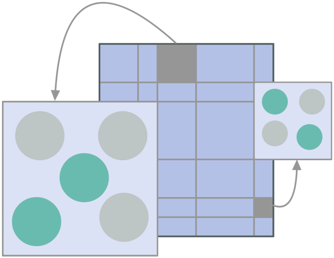
Complex sampling designs

Two-stage stratified cluster sampling



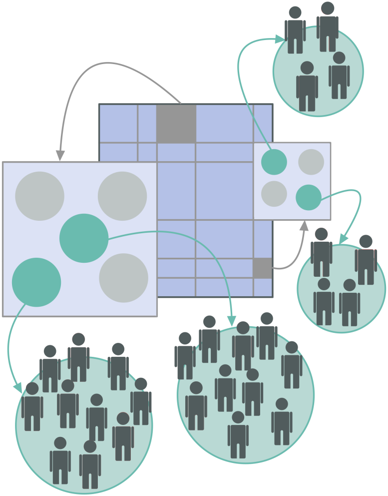
Complex sampling designs

Two-stage stratified cluster sampling



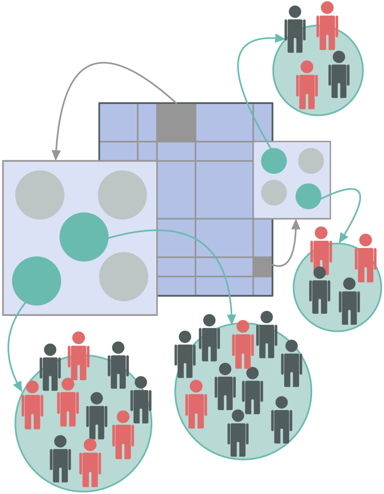
Complex sampling designs

Two-stage stratified cluster sampling



Complex sampling designs

Two-stage stratified cluster sampling



Complex sampling designs

Two-stage stratified cluster sampling

Population U (of size N):

$$U = \bigcup_{h=1}^{H} \bigcup_{\alpha=1}^{A_h} U_{h,\alpha} \text{ , each } U_{h,\alpha} \text{ of size } N_{h,\alpha}, \forall h \in \{1, \dots, H\}, \forall \alpha = 1, \dots, A_h.$$

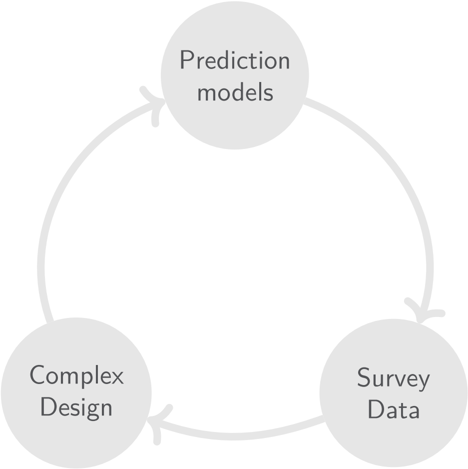
Inclusion probabilities:

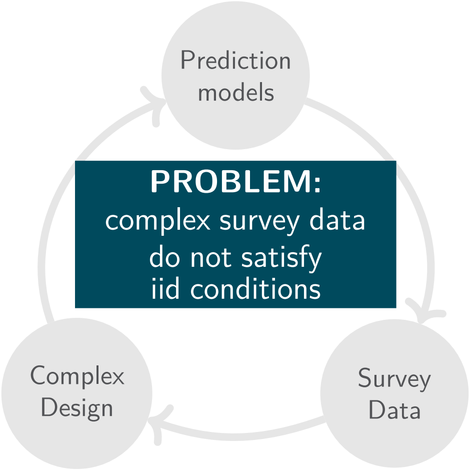
$$\pi_i = \frac{a_h}{A_h} \cdot \frac{n_{h,\alpha}}{N_{h,\alpha}}, \quad \forall i \in U_{h,\alpha}, \quad \forall \alpha \in \{1, \dots, A_h\}, \forall h \in \{1, \dots, H\}.$$

Sampling weights

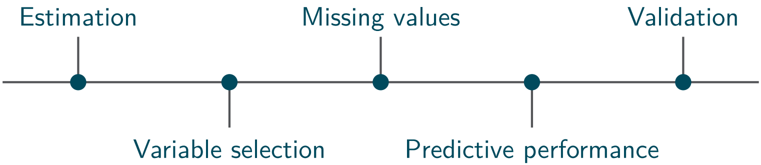
$$w_i = \frac{1}{\pi_i} = \frac{A_h}{a_h} \cdot \frac{N_{h,\dot{\alpha}}}{n_{h,\dot{\alpha}}}, \quad \forall i \in S_{h,\dot{\alpha}}, \forall \dot{\alpha} \in \mathbb{A}_h, \forall h \in \{1, \dots, H\},$$

where $\dot{\alpha}$ is the index of each selected cluster (grouped in the set $\mathbb{A}_h$).


Prediction models
Survey Data
Complex Design


Prediction models
PROBLEM:
complex survey data do not satisfy iid conditions
Survey Data
Complex Design

Objectives


Estimation
Missing values
Validation
Variable selection
Predictive performance

Objectives

Variable selection with LASSO regression for complex survey data

Objectives

Estimation of the ROC curve and AUC with complex survey data

Basic notation

Y: dichotomous response variable
$\boldsymbol{X} = (1, X_1, \ldots, X_p)$: vector of covariates.
U: finite population of N units
$S \subset U$: sample of n observations, $(y_i, \boldsymbol{x}_i, w_i), \forall i \in S$
$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$: model coefficients.

Focus: Logistic regression model

$$logit(p_i) = \ln\left[\frac{p_i}{1 - p_i}\right] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

where $p_i = p(\boldsymbol{x}_i) = P(Y = 1 | \boldsymbol{X} = \boldsymbol{x}_i)$.

Basic notation

Y: dichotomous response variable
$\boldsymbol{X} = (1, X_1, \ldots, X_p)$: vector of covariates.
U: finite population of N units
$S \subset U$: sample of n observations, $(y_i, \boldsymbol{x}_i, w_i), \forall i \in S$
$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$: model coefficients.

Focus: Logistic regression model

$$logit(p_i) = \ln\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

where $p_i = p(\boldsymbol{x}_i) = P(Y = 1 | \boldsymbol{X} = \boldsymbol{x}_i)$.

Likelihood function

$$L(\boldsymbol{\beta}) = \prod_{i \in S} p_i^{y_i}(1-p_i)^{1-y_i} \Longrightarrow \hat{\boldsymbol{\beta}}$$

Basic notation

Y: dichotomous response variable
$\boldsymbol{X} = (1, X_1, \ldots, X_p)$: vector of covariates.
U: finite population of N units
$S \subset U$: sample of n observations, $(y_i, \boldsymbol{x}_i, w_i), \forall i \in S$
$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$: model coefficients.

Focus: Logistic regression model

$$logit(p_i) = \ln\left[\frac{p_i}{1 - p_i}\right] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

where $p_i = p(\boldsymbol{x}_i) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x}_i)$.

Pseudo-likelihood function (Binder, 1983)

$$PL(\boldsymbol{\beta}) = \prod_{i \in S} p_i^{y_i w_i}(1 - p_i)^{(1-y_i)w_i} \Longrightarrow \hat{\boldsymbol{\beta}}$$

Variable selection with LASSO regression

Introduction

- Development of prediction models
  - Variable selection
  - LASSO regression models $\Longrightarrow$ Tuning parameter (λ)

Introduction

- Development of prediction models
  - Variable selection
  - LASSO regression models $\Longrightarrow$ Tuning parameter (λ)
  - Select λ that minimizes the error: validation methods (train/test sets)
  - Cross-validation (CV)

Introduction

- Development of prediction models
  - Variable selection
  - LASSO regression models $\Longrightarrow$ Tuning parameter (λ)
  - Select λ that minimizes the error: validation methods (train/test sets)
  - Cross-validation (CV)
- **PROBLEMS:** sampling design is not considered
  - Estimation of regression coefficients
  - Validation techniques

Introduction

- Development of prediction models
  - Variable selection
  - LASSO regression models $\Longrightarrow$ Tuning parameter (λ)
  - Select λ that minimizes the error: validation methods (train/test sets)
  - Cross-validation (CV)
- **PROBLEMS:** sampling design is not considered
  - Estimation of regression coefficients
  - Validation techniques
- Complex survey data framework:

Validation techniques $\Longrightarrow$ **Replicate weights methods**

Replicate weights methods

Modify the sampling weights (w_i^*) to define new subsamples that replicate the original sample and properly represent the finite population.

Introduction

Goals

1. Analyze the performance of replicate weights methods to select λ.
2. **Propose new methods** based on replicate weights: **design-based cross-validation (dCV)**.

Introduction

Goals

1. Analyze the performance of replicate weights methods to select λ.
2. **Propose new methods** based on replicate weights: **design-based cross-validation (dCV)**.

↓

We compare the performance of the methods with respect to the traditional cross-validation

Methods

- Logistic regression model: $p(\boldsymbol{x}_i) = \dfrac{e^{\boldsymbol{x}_i \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}}$

$$\ell(\boldsymbol{\beta}) = \sum_{i \in S} \left[y_i \ln(p(\boldsymbol{x}_i)) + (1 - y_i) \ln(1 - p(\boldsymbol{x}_i)) \right] \Longrightarrow \hat{\boldsymbol{\beta}}$$

- For a given value of λ, logistic LASSO regression models:

$$\min \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Methods

- Logistic regression model: $p(\boldsymbol{x}_i) = \dfrac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i\boldsymbol{\beta}}}$

$$\ell(\boldsymbol{\beta}) = \sum_{i \in S} \left[y_i \ln(p(\boldsymbol{x}_i)) + (1 - y_i) \ln(1 - p(\boldsymbol{x}_i)) \right] \Longrightarrow \hat{\boldsymbol{\beta}}$$

- For a given value of λ, logistic LASSO regression models:

$$\min \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \Longrightarrow \lambda?$$

Methods

Unweighted cross-validation (unw-SRSCV) $(K = 10)$

Methods

Unweighted cross-validation (unw-SRSCV) $(K = 10)$

> Define a grid for λ: λ_l, $\forall l = 1, \ldots, L$.

Methods

Unweighted cross-validation (unw-SRSCV) $(K = 10)$

- Define a grid for λ: λ_l, $\forall l = 1, \ldots, L$.
- K folds $\Longrightarrow S_{\text{tr}(k)},\ S_{\text{test}(k)}, \forall k = 1, \ldots, K$

Methods

Unweighted cross-validation (unw-SRSCV) $(K = 10)$

- Define a grid for λ: λ_l, $\forall l = 1, \ldots, L$.
- K folds $\Longrightarrow S_{\text{tr}(k)}, \; S_{\text{test}(k)}, \; \forall k = 1, \ldots, K$
- Fit the model to the training set $S_{\text{tr}(k)}$ considering $\lambda_l \Longrightarrow \hat{p}^{l}_{\text{tr}(k)}(\cdot)$

Methods

Unweighted cross-validation (unw-SRSCV) ($K = 10$)

- Define a grid for λ: λ_l, $\forall l = 1, \dots, L$.
- K folds $\Longrightarrow S_{\text{tr}(k)},\ S_{\text{test}(k)},\ \forall k = 1, \dots, K$
- Fit the model to the training set $S_{\text{tr}(k)}$ considering $\lambda_l \Longrightarrow \hat{p}^l_{\text{tr}(k)}(\cdot)$
- Estimate the error in the test sets:

$$\widehat{Err}^{\,l}_{(k)} = \frac{1}{n_{\text{test}(k)}} \sum_{i \in S_{\text{test}(k)}} \mathcal{L}(y_i, \hat{p}^l_{\text{tr}(k)}(\boldsymbol{x}_i)) \Longrightarrow \widehat{Err}_{CV}(\lambda_l) = \frac{1}{K} \sum_{k=1}^{K} \widehat{Err}^{\,l}_{(k)},$$

where: $\mathcal{L}(y_i, \hat{p}^l_{\text{tr}(k)}(\boldsymbol{x}_i)) = -y_i \ln(\hat{p}^l_{\text{tr}(k)}(\boldsymbol{x}_i)) - (1 - y_i) \ln(1 - \hat{p}^l_{\text{tr}(k)}(\boldsymbol{x}_i))$.

Methods

Unweighted cross-validation (unw-SRSCV) $(K = 10)$

- Define a grid for λ: λ_l, $\forall l = 1, \ldots, L$.
- K folds $\Longrightarrow S_{tr(k)},\ S_{test(k)},\ \forall k = 1, \ldots, K$
- Fit the model to the training set $S_{tr(k)}$ considering $\lambda_l \Longrightarrow \hat{p}^l_{tr(k)}(\cdot)$
- Estimate the error in the test sets:

$$\widehat{Err}^l_{(k)} = \frac{1}{n_{test(k)}} \sum_{i \in S_{test(k)}} \mathcal{L}(y_i, \hat{p}^l_{tr(k)}(\boldsymbol{x}_i)) \Longrightarrow \widehat{Err}_{CV}(\lambda_l) = \frac{1}{K} \sum_{k=1}^{K} \widehat{Err}^l_{(k)},$$

where: $\mathcal{L}(y_i, \hat{p}^l_{tr(k)}(\boldsymbol{x}_i)) = -y_i \ln(\hat{p}^l_{tr(k)}(\boldsymbol{x}_i)) - (1 - y_i) \ln(1 - \hat{p}^l_{tr(k)}(\boldsymbol{x}_i))$.

- Best value for λ:

$$\Lambda = \underset{\lambda_l:\ l=1,\ldots,L}{\operatorname{argmin}} \{\widehat{Err}_{CV}(\lambda_l)\}$$

Unweighted cross-validation (unw-SRSCV) ($K = 10$)

- Define a grid for λ: $\lambda_l, \forall l = 1, \ldots, L$.
- K folds $\Longrightarrow S_{tr(k)}, S_{test(k)}, \forall k = 1, \ldots, K$ (*)
- Fit the model to the training set $S_{tr(k)}$ considering $\lambda_l \Longrightarrow \hat{p}^l_{tr(k)}(\cdot)$ (*)
- Estimate the error in the test sets: (*)

$$\widehat{Err}^l_{(k)} = \frac{1}{n_{test(k)}} \sum_{i \in S_{test(k)}} \mathcal{L}(y_i, \hat{p}^l_{tr(k)}(\boldsymbol{x}_i)) \Longrightarrow \widehat{Err}_{CV}(\lambda_l) = \frac{1}{K} \sum_{k=1}^{K} \widehat{Err}^l_{(k)},$$

where:
$\mathcal{L}(y_i, \hat{p}^l_{tr(k)}(\boldsymbol{x}_i)) = -y_i \ln(\hat{p}^l_{tr(k)}(\boldsymbol{x}_i)) - (1 - y_i) \ln(1 - \hat{p}^l_{tr(k)}(\boldsymbol{x}_i))$.

- Best value for λ:

$$\Lambda = \underset{\lambda_l:\, l=1,\ldots,L}{\text{argmin}} \{\widehat{Err}_{CV}(\lambda_l)\}$$

Methods

PROPOSAL: Sampling design should be considered.

Methods

PROPOSAL: Sampling design should be considered.

- Fitting the models: $\Longrightarrow$ weighted cross-validation (w_i, w-SRSCV)

$$\min \left\{ -p\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where,

$$p\ell(\boldsymbol{\beta}) = \sum_{i \in S} w_i \left[y_i \ln(p(\boldsymbol{x}_i)) + (1 - y_i) \ln(1 - p(\boldsymbol{x}_i)) \right].$$

Methods

PROPOSAL: Sampling design should be considered.

- Fitting the models: $\Longrightarrow$ weighted cross-validation (w_i, w-SRSCV)

$$\min \left\{ -p\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where,

$$p\ell(\boldsymbol{\beta}) = \sum_{i \in S} w_i^* \left[y_i \ln(p(\boldsymbol{x}_i)) + (1 - y_i) \ln(1 - p(\boldsymbol{x}_i)) \right].$$

- Defining training and test sets $\Longrightarrow$ Replicate weights methods ($w_i{}^*$)

Methods

PROPOSAL: Sampling design should be considered.

- Fitting the models: $\Longrightarrow$ weighted cross-validation (w_i, w-SRSCV)

$$\min\left\{-p\ell(\beta)+\lambda\sum_{j=1}^{p}|\beta_j|\right\},$$

where,

$$p\ell(\beta)=\sum_{i\in S} w_i^*\left[y_i\ln(p(\mathbf{x}_i))+(1-y_i)\ln(1-p(\mathbf{x}_i))\right].$$

- Defining training and test sets $\Longrightarrow$ Replicate weights methods (${w_i}^*$)
- Estimating the error:

$$\widehat{Err}^{\,l}_{(k)}=\frac{1}{\sum_{i\in S_{\text{test}(k)}} w_i^*}\sum_{i\in S_{\text{test}(k)}} w_i^*\mathcal{L}(y_i,\hat{p}^{l}_{\text{tr}(k)}(\mathbf{x}_i)).$$

Methods

Replicate weights methods:

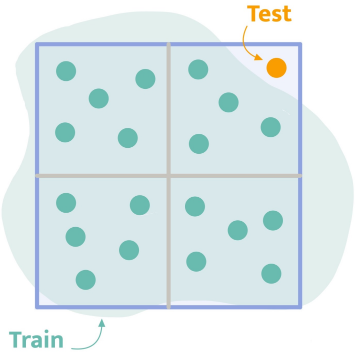
Existing methods	New methods
▶ Jackknife Repeated Replication (JKn)	▶ Design-based cross-validation (dCV)
▶ Rescaling Bootstrap (Bootstrap)	▶ Split-sample Repeated Replication (split)
▶ Balanced Repeated Replication (BRR)	▶ Extrapolation (extrap)

Methods

Replicate weights methods:

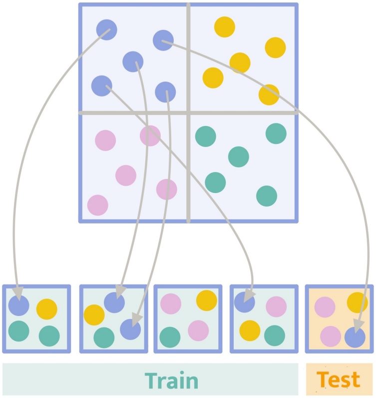
Existing methods	New methods
Jackknife Repeated Replication (JKn)	Design-based cross-validation (dCV)
Rescaling Bootstrap (Bootstrap)	Split-sample Repeated Replication (split)
Balanced Repeated Replication (BRR)	Extrapolation (extrap)

Jackknife Repeated Replication (JKn)



Methods

Design-based cross-validation (dCV)



Simulation study

- Generate population covariates ($\boldsymbol{x}_i$) and design variables ($\boldsymbol{z}_i$) following a multivariate normal distribution.
- Pre-define β (some values $= 0$) $\Longrightarrow y_i \sim Bernoulli(p(\boldsymbol{x}_i, \boldsymbol{z}_i)) \Longrightarrow U$

Simulation study

- Generate population covariates ($\boldsymbol{x}_i$) and design variables ($\boldsymbol{z}_i$) following a multivariate normal distribution.
- Pre-define β (some values $= 0$) $\Longrightarrow y_i \sim Bernoulli(p(\boldsymbol{x}_i, \boldsymbol{z}_i)) \Longrightarrow U$
- $p = 50$ covariates ($\boldsymbol{x}_i$) are considered to fit the models.

Simulation study

- Generate population covariates ($\boldsymbol{x}_i$) and design variables ($\boldsymbol{z}_i$) following a multivariate normal distribution.
- Pre-define β (some values $= 0$) $\Longrightarrow y_i \sim Bernoulli(p(\boldsymbol{x}_i, \boldsymbol{z}_i)) \Longrightarrow U$
- $p = 50$ covariates ($\boldsymbol{x}_i$) are considered to fit the models.
- S1 ($d = 0$ cluster-level variables), S2 ($d = 5$).

Simulation study

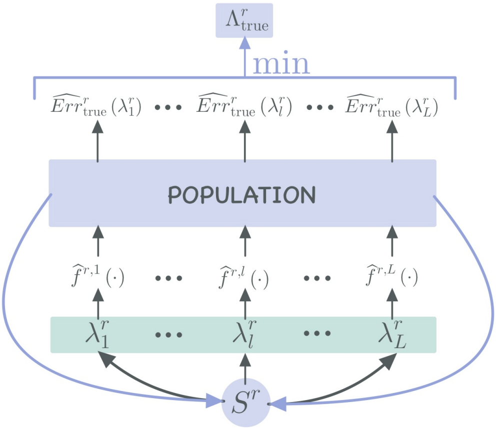
- Generate population covariates ($\boldsymbol{x}_i$) and design variables ($\boldsymbol{z}_i$) following a multivariate normal distribution.
- Pre-define β (some values $= 0$) $\Longrightarrow y_i \sim Bernoulli(p(\boldsymbol{x}_i, \boldsymbol{z}_i)) \Longrightarrow U$
- $p = 50$ covariates ($\boldsymbol{x}_i$) are considered to fit the models.
- S1 ($d = 0$ cluster-level variables), S2 ($d = 5$).
- $H = 5$ strata, $A_h = 20, \forall h = 1, \ldots, H$ clusters

Simulation study

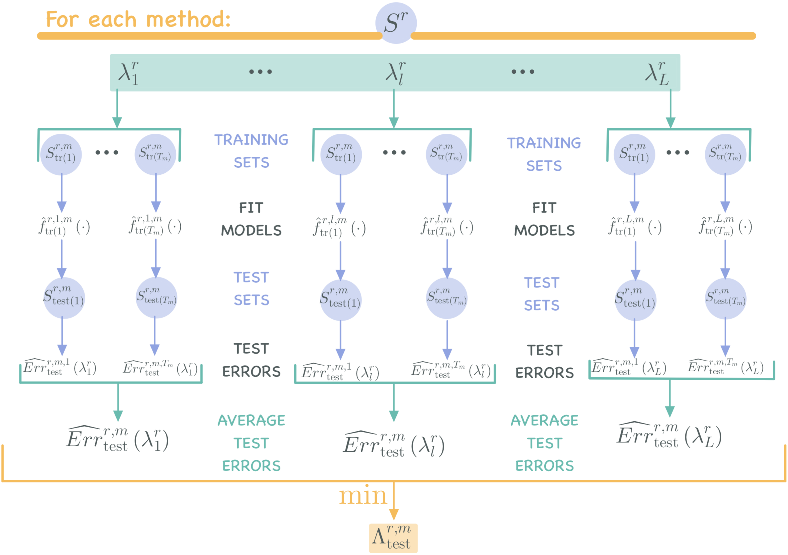
- Generate population covariates ($\boldsymbol{x}_i$) and design variables ($\boldsymbol{z}_i$) following a multivariate normal distribution.
- Pre-define β (some values $= 0$) $\Longrightarrow y_i \sim Bernoulli(p(\boldsymbol{x}_i, \boldsymbol{z}_i)) \Longrightarrow U$
- $p = 50$ covariates ($\boldsymbol{x}_i$) are considered to fit the models.
- S1 ($d = 0$ cluster-level variables), S2 ($d = 5$).
- $H = 5$ strata, $A_h = 20, \forall h = 1, \ldots, H$ clusters
- Sample (S):
  - $a_h = 4$, $\forall h = 1, \ldots, H$ clusters
  - $n_{h,\alpha}$ units per cluster:

    S1: $(5, 10, 25, 50, 500)$, **S2:** $(5, 25, 50, 100, 250) \Longrightarrow w_i$

Simulation study



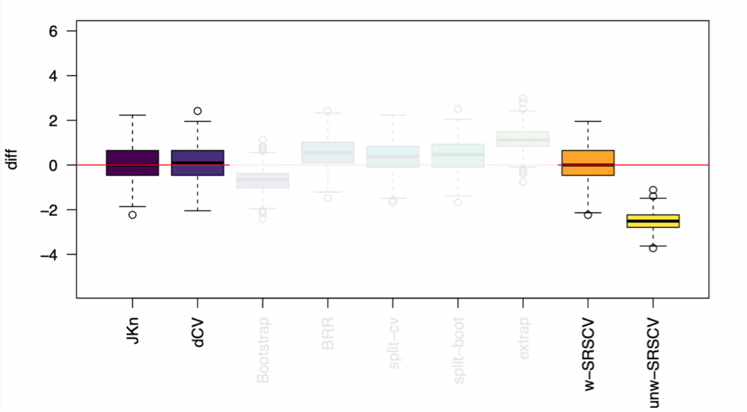
Simulation study



Simulation study

Differences between Λ parameters

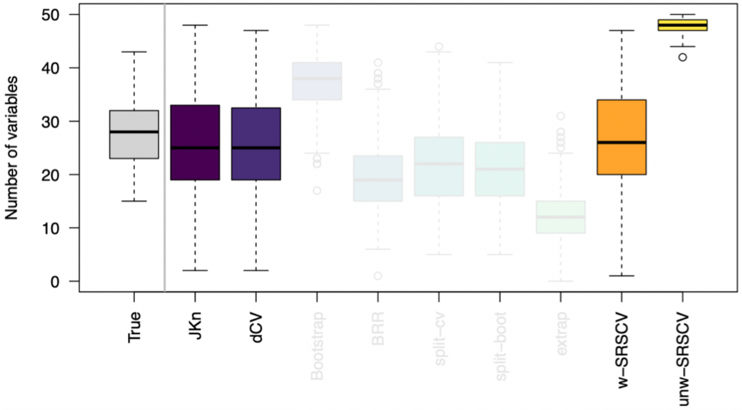
S1 ($d = 0$) **S2** ($d = 5$)



Simulation study

Number of variables

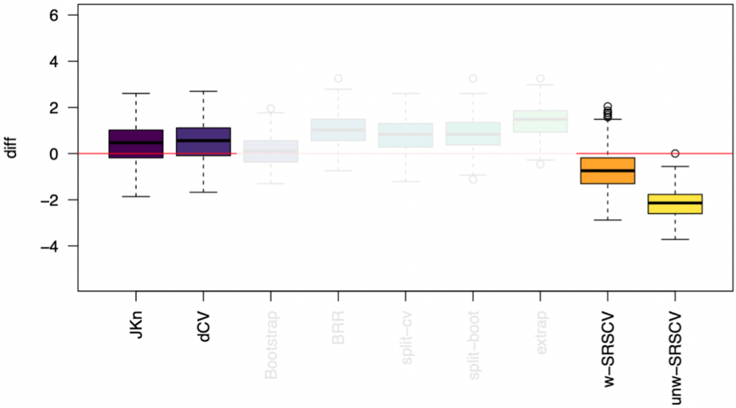
S1 ($d = 0$) **S2** ($d = 5$)



Simulation study

Differences between Λ parameters

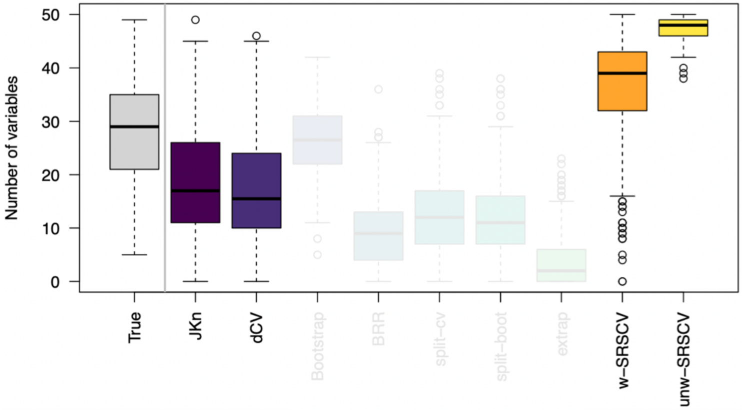
S1 ($d = 0$) **S2** ($d = 5$)



Simulation study

Number of variables

S1 ($d = 0$) S2 ($d = 5$)



Conclusions

- Weights need to be incorporated to fit LASSO models.
- The greater the cluster-effects, the greater the difference between dCV and w-SRSCV.
- Similar results for linear regression models.

Conclusions

- Weights need to be incorporated to fit LASSO models.
- The greater the cluster-effects, the greater the difference between dCV and w-SRSCV.
- Similar results for linear regression models.

Recommendation

The use of **dCV is recommended:** parsimonious models and the best method in terms of computational efficiency.

Conclusions

- Weights need to be incorporated to fit LASSO models.
- The greater the cluster-effects, the greater the difference between dCV and w-SRSCV.
- Similar results for linear regression models.

Recommendation

The use of **dCV is recommended:** parsimonious models and the best method in terms of computational efficiency.

Extended to elastic nets

Estimation of the ROC curve and the AUC

Introduction

S_0 : subset of units with $Y = 0$; S_1 : subset of units with $Y = 1$.

Introduction

S_0 : subset of units with $Y = 0$; S_1 : subset of units with $Y = 1$.

Area under the ROC curve ($\mathcal{A}_{\text{unw}}$)

$$\widehat{ROC}(\cdot) = \left\{(1 - \widehat{Sp}(c),\ \widehat{Se}(c)),\ c \in (-\infty, \infty)\right\} :$$

$$\widehat{Sp}(c) = \frac{1}{n} \sum_{i_0 \in S_0} I(\hat{p}_{i_0} < c) \quad ; \quad \widehat{Se}(c) = \frac{1}{n} \sum_{i_1 \in S_1} I(\hat{p}_{i_1} \geq c)$$

Introduction

S_0 : subset of units with $Y = 0$; S_1 : subset of units with $Y = 1$.

Area under the ROC curve ($\mathcal{A}_{\text{unw}}$)

$$\widehat{ROC}(\cdot) = \left\{(1 - \widehat{Sp}(c),\ \widehat{Se}(c)),\ c \in (-\infty, \infty)\right\}:$$

$$\widehat{Sp}(c) = \frac{1}{n} \sum_{i_0 \in S_0} I(\hat{p}_{i_0} < c) \quad ; \quad \widehat{Se}(c) = \frac{1}{n} \sum_{i_1 \in S_1} I(\hat{p}_{i_1} \geq c)$$

Mann-Whitney U-statistic (Bamber, 1975)

$$\widehat{AUC}_{\text{unw}} = \frac{1}{n_0 \cdot n_1} \sum_{i_0 \in S_0} \sum_{i_1 \in S_1} \left[I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 I(\hat{p}_{i_0} = \hat{p}_{i_1})\right]$$

$$\mathcal{A}_{\text{unw}} = \widehat{AUC}_{\text{unw}}$$

Proposal

S_0 : subset of units with $Y = 0$; S_1 : subset of units with $Y = 1$.

Area under the ROC curve ($\mathcal{A}$)

$$\widehat{ROC}_w(\cdot) = \left\{(1 - \widehat{Sp}_w(c),\ \widehat{Se}_w(c)),\ c \in (-\infty, \infty)\right\} :$$

$$\widehat{Sp}_w(c) = \frac{\sum_{i_0 \in S_0} w_{i_0} \cdot I(\hat{p}_{i_0} < c)}{\sum_{i_0 \in S_0} w_{i_0}} \quad ; \quad \widehat{Se}_w(c) = \frac{\sum_{i_1 \in S_1} w_{i_1} \cdot I(\hat{p}_{i_1} \geq c)}{\sum_{i_1 \in S_1} w_{i_1}}$$

Proposal

S_0 : subset of units with $Y = 0$; S_1 : subset of units with $Y = 1$.

Area under the ROC curve ($\mathcal{A}$)

$$\widehat{ROC}_w(\cdot) = \left\{(1 - \widehat{Sp}_w(c), \widehat{Se}_w(c)), \ c \in (-\infty, \infty)\right\} :$$

$$\widehat{Sp}_w(c) = \frac{\sum_{i_0 \in S_0} w_{i_0} \cdot I(\hat{p}_{i_0} < c)}{\sum_{i_0 \in S_0} w_{i_0}} \quad ; \quad \widehat{Se}_w(c) = \frac{\sum_{i_1 \in S_1} w_{i_1} \cdot I(\hat{p}_{i_1} \geq c)}{\sum_{i_1 \in S_1} w_{i_1}}$$

Based on the Mann-Whitney U-statistic

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} \left[I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})\right]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$

Proposal

S_0 : subset of units with $Y = 0$; S_1 : subset of units with $Y = 1$.

Area under the ROC curve ($\mathcal{A}$)

$$\widehat{ROC}_w(\cdot) = \left\{(1 - \widehat{Sp}_w(c),\ \widehat{Se}_w(c)),\ c \in (-\infty, \infty)\right\} :$$

$$\widehat{Sp}_w(c) = \frac{\sum_{i_0 \in S_0} w_{i_0} \cdot I(\hat{p}_{i_0} < c)}{\sum_{i_0 \in S_0} w_{i_0}} \quad ; \quad \widehat{Se}_w(c) = \frac{\sum_{i_1 \in S_1} w_{i_1} \cdot I(\hat{p}_{i_1} \geq c)}{\sum_{i_1 \in S_1} w_{i_1}}$$

Based on the Mann-Whitney U-statistic

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} \left[I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})\right]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$

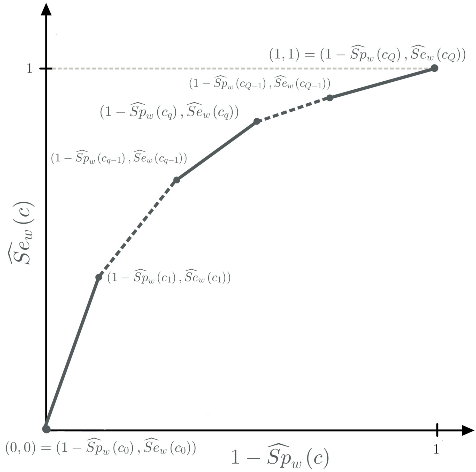
$$\mathcal{A} = \widehat{AUC}_w$$

Proposal

Q probabilities $\Longrightarrow$ Cut-off points: $c_Q < c_{Q-1} < \ldots < c_1 < c_0$

Proposal

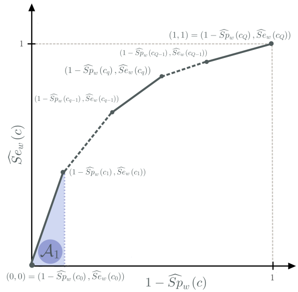
Q probabilities $\Longrightarrow$ Cut-off points: $c_Q < c_{Q-1} < \ldots < c_1 < c_0$
$\Longrightarrow \forall q \in \{0, 1, \ldots, Q\}, \quad (1 - \widehat{Sp}_w(c_q), \widehat{Se}_w(c_q)) \Longrightarrow \widehat{ROC}_w(\cdot)$



Proposal

Q probabilities $\Longrightarrow$ Cut-off points: $c_Q < c_{Q-1} < \ldots < c_1 < c_0$

$\Longrightarrow \forall q \in \{0, 1, \ldots, Q\}, \quad (1 - \widehat{Sp}_w(c_q), \widehat{Se}_w(c_q)) \Longrightarrow \widehat{ROC}_w(\cdot)$

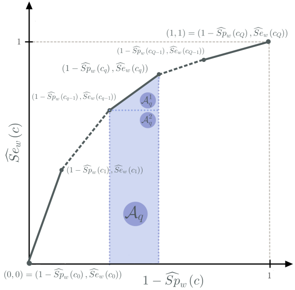


$$\mathcal{A}_1 = \frac{(1 - \widehat{Sp}_w(c_1)) \cdot \widehat{Se}_w(c_1)}{2}$$

Proposal

Q probabilities $\Longrightarrow$ Cut-off points: $c_Q < c_{Q-1} < \ldots < c_1 < c_0$

$\Longrightarrow \forall q \in \{0, 1, \ldots, Q\}, \quad (1 - \widehat{Sp}_w(c_q), \widehat{Se}_w(c_q)) \Longrightarrow \widehat{ROC}_w(\cdot)$



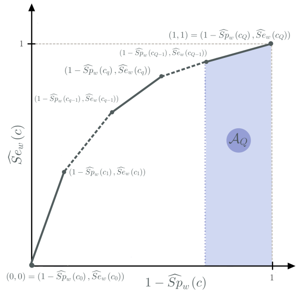
$$\mathcal{A}_1 = \frac{(1 - \widehat{Sp}_w(c_1)) \cdot \widehat{Se}_w(c_1)}{2}$$

$$\vdots$$

$$\mathcal{A}_q = \frac{(\widehat{Sp}_w(c_{q-1}) - \widehat{Sp}_w(c_q)) \cdot (\widehat{Se}_w(c_q) + \widehat{Se}_w(c_{q-1}))}{2}$$

Proposal

Q probabilities $\Longrightarrow$ Cut-off points: $c_Q < c_{Q-1} < \ldots < c_1 < c_0$
$\Longrightarrow \forall q \in \{0, 1, \ldots, Q\}, \quad (1 - \widehat{Sp}_w(c_q), \widehat{Se}_w(c_q)) \Longrightarrow \widehat{ROC}_w(\cdot)$



$$\mathcal{A}_1 = \frac{[1 - \widehat{Sp}_w(c_1)] \cdot \widehat{Se}_w(c_1)}{2}$$

$$\vdots$$

$$\mathcal{A}_q = \frac{[\widehat{Sp}_w(c_{q-1}) - \widehat{Sp}_w(c_q)] \cdot [\widehat{Se}_w(c_q) + \widehat{Se}_w(c_{q-1})]}{2}$$

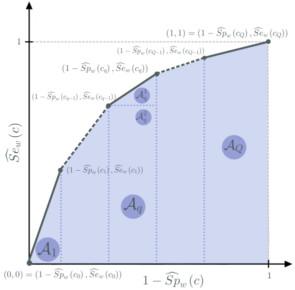
$$\vdots$$

$$\mathcal{A}_Q = \frac{\widehat{Sp}_w(c_{Q-1}) \cdot [1 + \widehat{Se}_w(c_{Q-1})]}{2}$$

Proposal

Q probabilities $\Longrightarrow$ Cut-off points: $c_Q < c_{Q-1} < \ldots < c_1 < c_0$

$\Longrightarrow \forall q \in \{0, 1, \ldots, Q\}, \quad (1 - \widehat{Sp}_w(c_q), \widehat{Se}_w(c_q)) \Longrightarrow \widehat{ROC}_w(\cdot)$



$$\mathcal{A} = \mathcal{A}_1 + \ldots + \mathcal{A}_Q$$

$$\Downarrow$$

$$\mathcal{A} = \frac{1}{2}\sum_{q=1}^{Q}[\widehat{Sp}_w(c_{q-1})\widehat{Se}_w(c_q) - \widehat{Sp}_w(c_q)\widehat{Se}_w(c_{q-1})]$$

Proposal

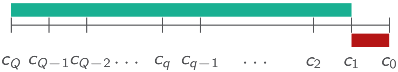
$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} \left[I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1}) \right]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$

Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} \left[I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1}) \right]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$

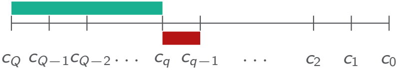
Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} [I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$



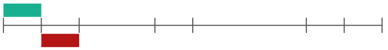
Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} [I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$



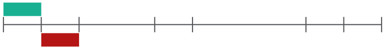
Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} \left[I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1}) \right]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$



Proposal

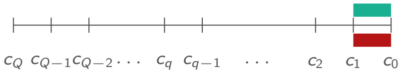
$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} \left[I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1}) \right]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$



$$I(\hat{p}_{i_0} < \hat{p}_{i_1}) = \sum_{q=1}^{Q} I(\hat{p}_{i_0} < c_q) \cdot \left[I(\hat{p}_{i_1} \geq c_q) - I(\hat{p}_{i_1} \geq c_{q-1}) \right]$$

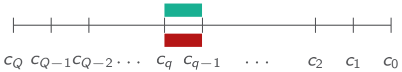
Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} [I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$



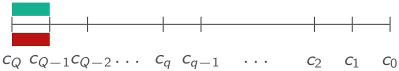
Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} [I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$



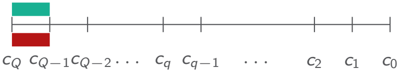
Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} [I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$



Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} [I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$



$$I(\hat{p}_{i_0} = \hat{p}_{i_1}) = \sum_{q=1}^{Q} [I(\hat{p}_{i_0} < c_{q-1}) - I(\hat{p}_{i_0} < c_q)] \cdot [I(\hat{p}_{i_1} \geq c_q) - I(\hat{p}_{i_1} \geq c_{q-1})]$$

Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} [I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1})]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$

$$I(\hat{p}_{i_0} < \hat{p}_{i_1}) = \sum_{q=1}^{Q} I(\hat{p}_{i_0} < c_q) \cdot [I(\hat{p}_{i_1} \geq c_q) - I(\hat{p}_{i_1} \geq c_{q-1})]$$

$$I(\hat{p}_{i_0} = \hat{p}_{i_1}) = \sum_{q=1}^{Q} [I(\hat{p}_{i_0} < c_{q-1}) - I(\hat{p}_{i_0} < c_q)] \cdot [I(\hat{p}_{i_1} \geq c_q) - I(\hat{p}_{i_1} \geq c_{q-1})]$$

Proposal

$$\widehat{AUC}_w = \frac{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1} \left[I(\hat{p}_{i_0} < \hat{p}_{i_1}) + 0.5 \cdot I(\hat{p}_{i_0} = \hat{p}_{i_1}) \right]}{\sum_{i_0 \in S_0} \sum_{i_1 \in S_1} w_{i_0} w_{i_1}}$$

$$=$$

$$\mathcal{A} = \frac{1}{2} \sum_{q=1}^{Q} [\widehat{Sp}_w(c_{q-1}) \widehat{Se}_w(c_q) - \widehat{Sp}_w(c_q) \widehat{Se}_w(c_{q-1})]$$

Simulation study

Data generation

Step 1. Generate U with covariates following a normal distribution.

Simulation study

Data generation

Step 1. Generate U with covariates following a normal distribution.

Step 2. Generate the response for a given β following Bernoulli's distribution.

Simulation study

Data generation

Step 1. Generate U with covariates following a normal distribution.

Step 2. Generate the response for a given β following Bernoulli's distribution.

Step 3. Define the sampling design:

- Strata
- Clusters within strata

Simulation study

Data generation

Step 1. Generate U with covariates following a normal distribution.

Step 2. Generate the response for a given β following Bernoulli's distribution.

Step 3. Define the sampling design:

- Strata
- Clusters within strata

Step 4. Sample the population and calculate the weights:

Simulation study

Data generation

Step 1. Generate U with covariates following a normal distribution.

Step 2. Generate the response for a given β following Bernoulli's distribution.

Step 3. Define the sampling design:

- Strata
- Clusters within strata

Step 4. Sample the population and calculate the weights:

- One-stage stratified sampling (SH)
- Two-stage stratified cluster sampling (SC)
  - 0 cluster-level variables (SC.0)
  - 1 cluster-level variable (SC.1)
- Two sampling schemes: (a) and (b)

Simulation study

Data generation

Step 1. Generate U with covariates following a normal distribution.

Step 2. Generate the response for a given β following Bernoulli's distribution.

Step 3. Define the sampling design:

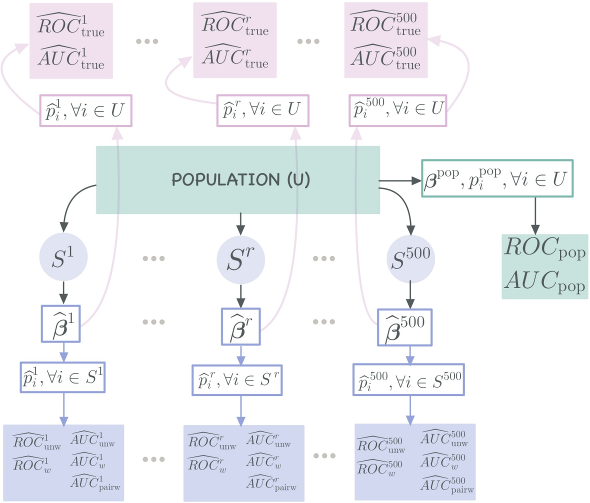
- Strata
- Clusters within strata

Step 4. Sample the population and calculate the weights:

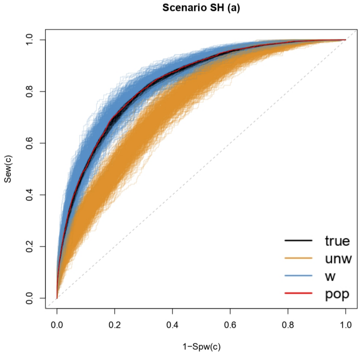
- One-stage stratified sampling (SH)
- Two-stage stratified cluster sampling (SC)
  - 0 cluster-level variables (SC.0)
  - 1 cluster-level variable (SC.1)
- Two sampling schemes: (a) and (b)

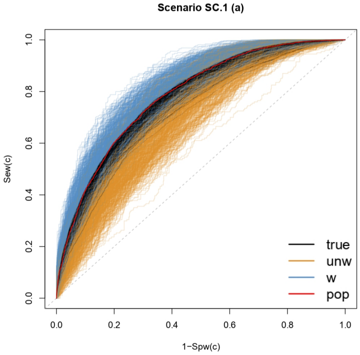
Simulation study

Set-up

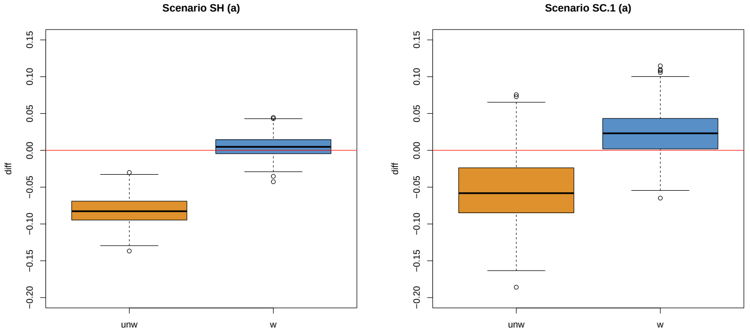


Estimated ROC curves





Differences between the estimated and true AUCs



Problem

$$\widehat{AUC}_w$$

$$\Downarrow$$

Optimistic estimates

$$\Downarrow$$

Correction needed

Optimism correction of the AUC

Introduction

Same data: (1) fit the model, (2) estimate the AUC $\Longrightarrow$ **Optimism**

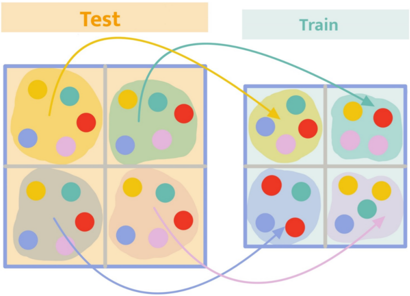
- In line with traditional simple random sample (SRS) context
  See, e.g: Austin and Steyerberg (2017), Iparragirre et al (2019).
- Recommendation in SRS: validation techniques
  - split-sample validation
  - Bootstrap
  - cross-validation
- In general, in complex survey data context, to define training and test sets
  - Validation techniques $\Longrightarrow$ **Replicate weights**

Goal

Analyze the performance of replicate weights methods for optimism correction of the AUC.

Replicate weights

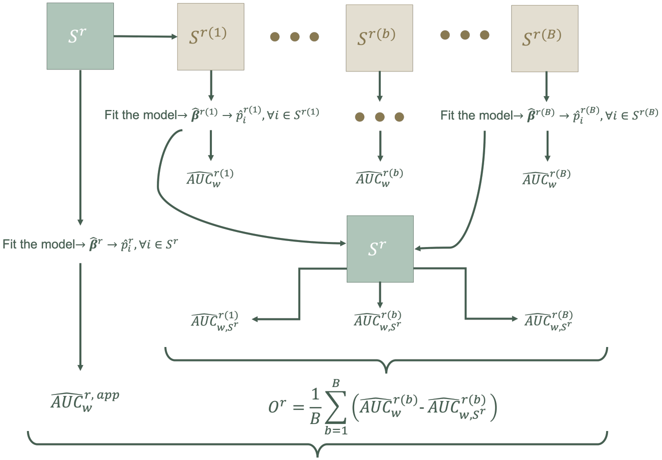
Rescaling Bootstrap (RB) (Rao and Wu, 1988)



RBn: another variant in which the same number of units (one-stage) or clusters (two-stage) are in both, training and test (original) set.

Replicate weights

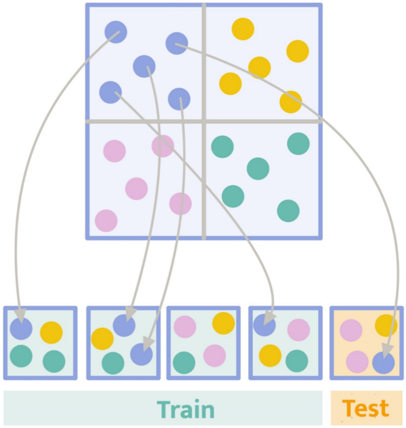
Rescaling Bootstrap (RB)



$$\widehat{AUC}_w^{r,RB} = \widehat{AUC}_w^{r,app} - O^r$$

Replicate weights

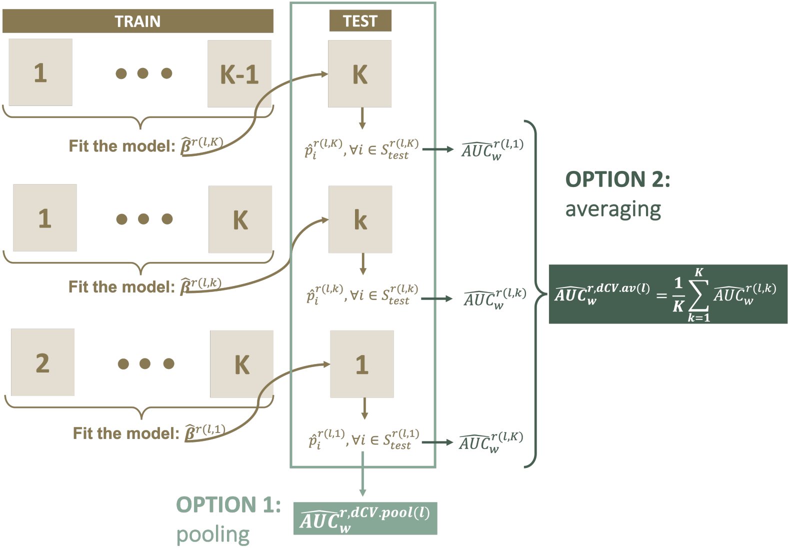
Design-based cross-validation (dCV) (Iparragirre et al. (2023))



JKn: another variant in which each unit (one-stage) or cluster (two-stage) is set as the test set once.

Replicate weights

Design-based cross-validation (dCV)

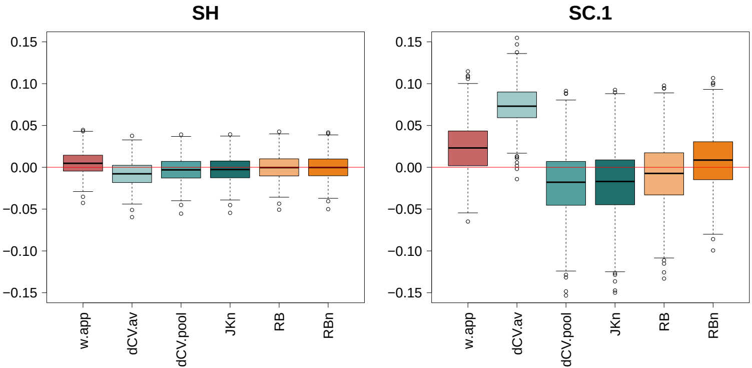


Simulation study

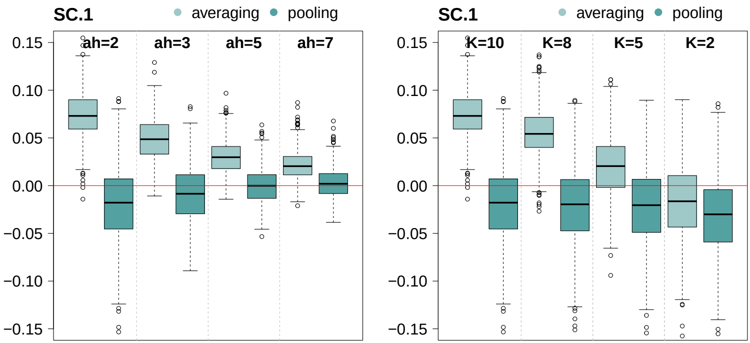
- Population U is generated in the same way as in the previous simulation study carried out for the estimation of the ROC curve.
- Sampling schemes: SH (one-stage), and SC.0, SC.1 (two-stage).
- Consider:
  - RB, RBn: $B = 200$ resamples
  - dCV: $K = 10$ folds, $L = 20$ replicates
- **Simulation set-up:** For $r = 1, \ldots, 500$:
  - Obtain the sample S^r
  - Fit the model to S^r ($\hat{\beta}^r$) and estimate its AUC: $\widehat{AUC}_w^{r,\text{app}}$.
  - Calculated the corrected AUCs : dCV.av, dCV.pool, JKn, RB, RBn
  - Extend $\hat{\beta}^r$ to U: $\widehat{AUC}_{\text{true}}^r$
  - For $m \in \{\text{app, dCV.av, dCV.pool, JKn, RB, RBn}\}$:, $\widehat{AUC}_w^{r,m}$

$$\text{diff}^{\,r,m} = \widehat{AUC}_w^{r,m} - \widehat{AUC}_{\text{true}}^r$$

Simulation study



Simulation study



Conclusions

Conclusions

- We propose unbiased design-based estimators for estimating the ROC curve and AUC in the context of complex survey data.
- Replicate weights recommended for the optimism correction of the AUC.

Conclusions

Conclusions

- We propose unbiased design-based estimators for estimating the ROC curve and AUC in the context of complex survey data.
- Replicate weights recommended for the optimism correction of the AUC.

Further research

- Variance estimation and confidence intervals for the ROC curve and AUC
- Extended simulation study to properly understand the behaviour of each replicate weight methods under different scenarios for optimism correction.

`svyVarSel` R package

svyVarSel

Goal

Variable selection with complex survey data.

- Initially: LASSO regression models $\Longrightarrow$ Extended to: Elastic Nets

Available functions:

Function	Brief description
`replicate.weights()`	Define training and test sets with replicate weights.
`wlasso()`	Fit LASSO models with complex survey data.
`welnet()`	Fit elastic nets with complex survey data.
`wlasso.plot()`	Graphical visualization of the error.
`welnet.plot()`	Graphical visualization of the error.

CRAN The Comprehensive R Archive Network

https://github.com/aiparragirre/svyVarSel
https://cran.r-project.org/web/packages/svyVarSel/

svyVarSel: welnet()

Purpose

Fit elastic net models with complex survey data.

Formulation:

$$\min \left\{ -p\ell(\boldsymbol{\beta}) + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right\} \Longrightarrow \lambda?$$

Steps: For a grid of values for λ_k, $k \in \{1, \ldots, K\}$,

1. Define train and test sets
2. Fit the models in the train set
3. Estimate the error of the fitted model in the test set

Select: $\lambda_k \in \{\lambda_1, \ldots, \lambda_K\}$, that minimizes the error

svyVarSel: welnet()

Usage

```
mcv <- welnet(data = simdata_lasso_binomial,
              col.y = "y", col.x = 1:50,
              family = "binomial",
              alpha = 0.5,
              cluster = "cluster", strata = "strata", weights = "weights",
              method = "dCV", k=10, R=20)
```

`svyVarSel: welnet()`

Output

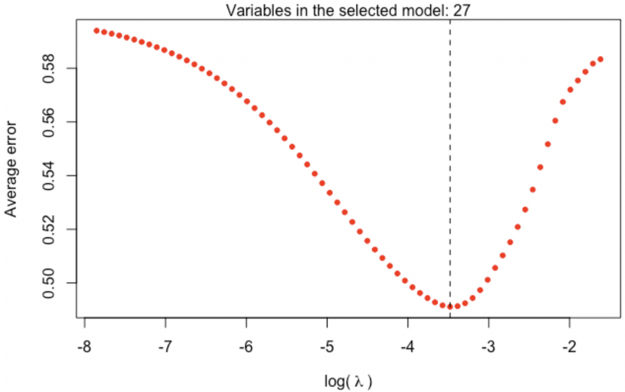
A list containing the following elements:

- `lambda`:
  - `grid`: All the values in the grid $\{\lambda_1, \ldots, \lambda_K\}$.
  - `min`: The value of $\lambda_k \in \{\lambda_1, \ldots, \lambda_K\}$ that minimizes the error.
- `error`:
  - `average`: average error for each $\lambda_k \in \{\lambda_1, \ldots, \lambda_K\}$
  - `all`: error for each $\lambda_k \in \{\lambda_1, \ldots, \lambda_K\}$ in each test set.
- `model`:
  - `grid`: all the coefficients of all the fitted models for $\{\lambda_1, \ldots, \lambda_K\}$.
  - `min`: model coefficients considering the λ_k that minimizes the error.
- `data.rw`:
  - Data frame with the information of the training and test sets defined with replicate weights.

svyVarSel: welnet.plot()

Usage and output

```
welnet.plot(mcv)
```



svyROC R package

svyROC

Goal

Estimation of the ROC curve, AUC and optimal cut-off points with complex survey data.

Available functions:

Function	Brief description
`wsp()`, `wse()`	Estimate the specificity and sensitivity parameters
`wocp()`	Estimate optimal cut-off points
`wauc()`	Estimate the AUC
`corrected.wauc()`	Corrected estimate of the AUC based on replicate weights
`wroc()`	Estimate the ROC curve
`wroc.plot()`	Plot the ROC curve

CRAN The Comprehensive R Archive Network

https://github.com/aiparragirre/svyROC
https://cran.r-project.org/web/packages/svyROC/

svyROC: wroc()

Usage

```
mycurve <- wroc(response.var = "y",
                phat.var = "phat",
                weights.var = "weights",
                data = example_data_wroc,
                tag.event = 1,
                tag.nonevent = 0,
                cutoff.method = "Youden")
```

svyROC: wroc()

Output

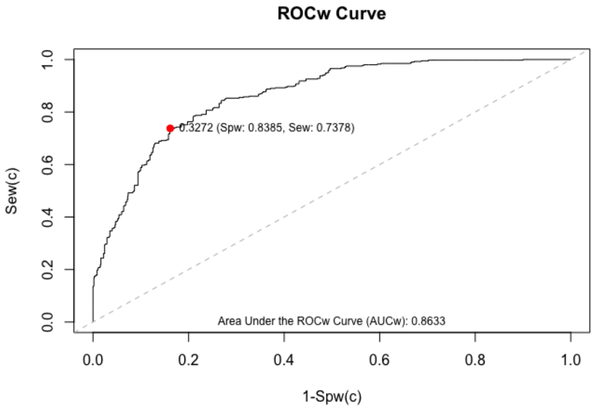
A list containing the following elements:

- `wroc.curve`: list containing the following elements:
  - `Sew.values`, `Spw.values`: all the values of the weighted estimate of sensitivity and specificity across all the possible cut-off points.
  - `cutoffs`: all the evaluated cut-off points.
- `wauc`: a numeric value indicating the area under the curve.
- `optimal.cutoff`: list containing the following elements:
  - `method`: Youden, ROC01, MaxProdSpSe or MaxEfficiency
  - `cutoff.value`: optimal cut-off point
  - `Spw, Sew`: sensitivity and specificity estimates for the optimal cutoff
- Other basic information

svyROC: wroc.plot()

Usage and output

```
wroc.plot(x = mycurve,
          print.auc = TRUE,
          print.cutoff = TRUE)
```



svyROC: corrected.wauc()

Usage

```
cor <- corrected.wauc(data = example_variables_wroc,
                      formula = y ~ x1 + x2 + x3 + x4 + x5 + x6,
                      tag.event = 1, tag.nonevent = 0,
                      weights.var = "weights", strata.var = "strata", cluster.var = "cluster",
                      method = "dCV", dCV.method = "pooling", k = 10, R = 20)
```

Output

A list containing:

- `corrected.AUCw`: the value of the corrected AUC.
- Other basic information

Discussion and further research

New proposals improve the development of prediction models

- Variable selection based on elastic nets
  - Design-based cross-validation
- Unbiased estimators for the ROC curve and AUC
  - Optimism correction based on replicate weights
- **Easy to apply:** implemented in `svyVarSel` and `svyROC`

Discussion and further research

New proposals improve the development of prediction models

- Variable selection based on elastic nets
  - Design-based cross-validation
- Unbiased estimators for the ROC curve and AUC
  - Optimism correction based on replicate weights
- **Easy to apply:** implemented in `svyVarSel` and `svyROC`

Further research

- Variable selection with **Statistical Boosting** for complex survey data.
- **Variance estimation** and confidence intervals for the ROC and AUC.
- Implement the proposals in `svyVarSel` and `svyROC`.

References

More methodological details

- Variable selection

Iparragirre, A., Lumley, T., Barrio, I., & Arostegui, I. (2023).
Variable selection with LASSO regression for complex survey data.
Stat, 12(1), e578.

- ROC curve and AUC

Iparragirre, A., Barrio, I., & Arostegui, I. (2023).
Estimation of the ROC curve and the area under it with complex survey data.
Stat, 12(1), e635.

Iparragirre, A., Barrio, I., Aramendi, J. & Arostegui, I. (2022).
Estimation of cut-off points under complex-sampling design data.
SORT-Statistics and Operations Research Transactions, 46(1), 137–158.

Iparragirre, A., Barrio, I. (2024).
Optimism Correction of the AUC with Complex Survey Data.
In: Einbeck, J., Maeng, H., Ogundimu, E., Perrakis, K. (eds) *Developments in Statistical Modelling. IWSM 2024. Contributions to Statistics*. Springer, Cham.

References

Related main references

Bamber, D.
The area above the ordinal dominance graph and the area below the receiver operating characteristic graph.
Journal of Mathematical Psychology, 1975; 12(4), 387-415.

Binder, D. A.
On the variances of asymptotically normal estimators from complex surveys.
International Statistical Review/Revue Internationale de Statistique, 1983; 279-292.

Iparragirre, A., Barrio, I., & Rodríguez-Álvarez, M. X.
On the optimism correction of the area under the receiver operating characteristic curve in logistic prediction models.
SORT - Statistics and Operations Research Transactions, 2019; 43(1), 145–162.

Rao, J. N. K., & Wu, C. F. J.
Resampling Inference With Complex Survey Data.
Journal of the American Statistical Association, 1988; 83(401), 231–241.

Tsuruta, H., & Bax, L.
Polychotomization of continuous variables in regression models based on the overall C index.
BMC Medical Informatics and Decision Making, 2006; 6(1), 41.

Yao, W., Li, Z., & Graubard, B. I.
Estimation of ROC curve with complex survey data.
Statistics in Medicine, 2015; 34(8), 1293-1303.

Thank you for your attention