05. General Model diagnosis

For classic models like linear regression models and GLM, the goodness of fit and predictability can be assessed by:

- various hypothesis test (e.g. Wald test, Score test, and likelihood ratio test); and. → training error.
- model selection criteria (e.g. AIC, BIC and various IC) → make adjustment to training error to estimate testing error indirectly.

General model assessment methods:

(1) resampling: CV → estimate testing error directly (more in 03); bootraps (statistical inference like CI)

(2) Gains & ROC → better if applied to predictions made on a validation set. (to help model selection).

(3) Actual versus Expected prediction (fitted) plots

(4) Partial dependence plots

First we need to split the data into:

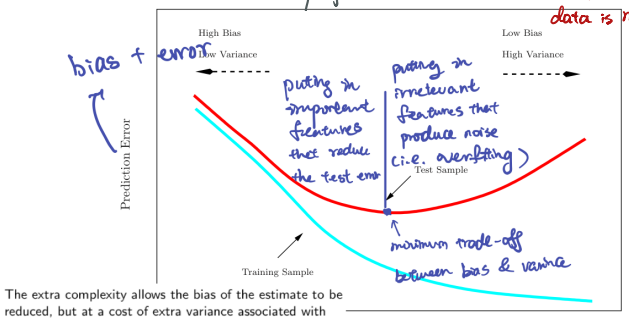
TRAINING SET	DEV/VALIDATION SET	TESTING SET
↓	↓	↓
model fitting	model selecting (e.g. select model from subsets select, or selecting λ for lasso)	model assessment. (i.e. prediction accuracy).

* usually split into 2:1:1 for small dataset (<10,000)

* usually split into 98:1:1 for large dataset (1,000,000)

Why do we use test data to assess model performance:

- TEST error is the average error that results from using a model to predict the response on new observations
- TRAN error is the residuals that result from a model fitting on observations
- If we use TRAN error, we tend to over-fit the data, and include unnecessary terms in the model; TRAN error doesn't tell us how well the model perform on new data. => a model that fits well on training data is not indicative of true performance.



bias + error ↑

putting in important features that reduce the test error

putting in irrelevant features that produce noise (i.e. overfitting)

minimum trade-off between bias & variance

The extra complexity allows the bias of the estimate to be reduced, but at a cost of extra variance associated with estimating parameters.

* model complexity ↑, Tr error ↓

Te error ↑ (due to overfitting)

Model Complexity → e.g. # feature from low to high, degree of polynomial from low to high.

(2) Gains curve & ROC curve.

2.1 Gains Curve

method: low predictions should correspond to low response observations, and high predictions should correspond to high response observations

=> order the observations based on predicted response descendingly, and see how well the actual responses are sorted descendingly.

=> then include the cumulative sums of observations, as well as the %cumsum/total sum.

=> better model will have larger cumsum earlier since the larger predicted responses appear earlier.

steps: best way to describe is to use an example: suppose we have 2 candidates f_1, f_2 to predict Y:

obsern. model1 model2

Y_i	$f_1(X_i)$	$f_2(X_i)$
2.0	2.0	3.2
2.5	2.5	2.5
3.0	3.0	3.0
3.0	4.0	4.0
4.5	5.0	3.5
5.0	4.5	4.5
5.5	6.0	3.4

sorting Y_i in concordant with $f_1(x_i)$

$f_1(X_i)$	Y_i
6.0	5.5
5.0	4.5
4.5	5.0
4.0	3.0
3.0	3.0
2.5	2.5
2.0	2.0

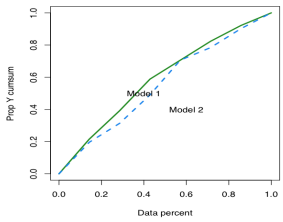
$f_2(X_i)$	Y_i
4.5	5.0
4.0	3.0
3.5	4.5
3.4	5.5
3.2	2.0
3.0	3.0
2.5	2.5

Cumsum

$f_1(X_i)$	Y_i	CumSum	Percent
6.0	5.5	5.5	0.22 = 5.5/25.5
5.0	4.5	10.0 (+4.5)	0.39 = 10/25.5
4.5	5.0	15.0 (+5)	0.59
4.0	3.0	18.0 ⋮	0.71
3.0	3.0	21.0 ⋮	0.82
2.5	2.5	23.5	0.92
2.0	2.0	25.5 (+2)	1.00

$f_2(X_i)$	Y_i	CumSum	Percent
4.5	5.0	5.0	0.20
4.0	3.0	8.0	0.31
3.5	4.5	12.5	0.49
3.4	5.5	18.0	0.71
3.2	2.0	20.0	0.78
3.0	3.0	23.0	0.90
2.5	2.5	25.5	1.00

plot against $\frac{j}{n}$



⇒ model 1 is better

mathematic representations:

Gain curve plots $\frac{\sum_{i=0}^{j} Y_{k_i}}{\sum_{k=1}^{n} Y_k}$ versus $\frac{j}{n}$, $j=0,1,\dots n$

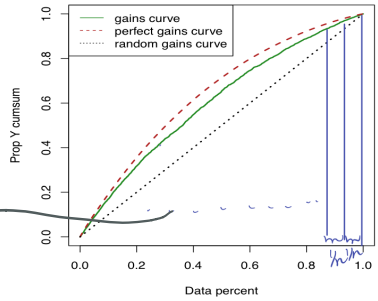
where $k_1, k_2, \dots k_n$ are ordered index of $f(X_{k_1}) \geq f(X_{k_2}) \dots \geq f(X_{k_n})$ ⇒ $f(X_{k_n}) = f(X)_{(n)}$

Thus, the points can be represented as: $(G_X(t), G_Y(t))$ where

$$G_X(t) = \frac{\sum_{i=1}^{n} I(f(x_i) > t)}{n}, \quad G_Y(t) = \frac{\sum_{i=1}^{n} Y_i I(f(x_i) \geq t)}{\sum_{i=1}^{n} Y_i} \quad \text{where } t \in [\min_{1\leq i\leq n} Y_i, \max_{1\leq i\leq n} Y_i].$$

Compare with 'perfect curve' and 'random curve':

- perfect curve: prediction is observation itself
- random curve: randomly prediction i.e. 45° line



AUC (area under curve):

$$AUC = \left(2n \sum_{k=1}^{n} Y_{(k)}\right)^{-1} \sum_{j=0}^{n-1} \left[\sum_{i=0}^{j} Y_{k_i} + \sum_{i=0}^{j+1} Y_{k_i}\right]$$

problem: only compares models in terms of how well the ordering of observations is concordant with that of their prediction ⇒ not able to assess goodness of fit.

2.2 ROC (receiver operating characteristic) (similar to Gains curve but Y is binary)

method: First construct a confusion matrix; then plot sensitivity rate against false positive rate also by different cut-off value t; we prefer models with larger tp and tn, and smaller fp and fn.

steps:

(1) construct:

Confusion matrix:

	prediction		
Actual	class 0	class 1	total
class 0	tn	fp	tn+fp (useful)
class 1	fn	tp	fn+tp (useful)
Total	tn+fn (not useful)	fp+tp (not useful)	n

⇒ 1=

- sensitivity rate (true positive rate) $= \frac{tp}{tp+fn}$ = 1 if cut-off value = 0 (% of real 1s predicted as 1)
- specificity rate (true negative rate) $= \frac{tn}{tn+fp}$ = 1 if cut-off value = 1 (% of real 0s predicted as 0)
- false positive rate $= \frac{fp}{tn+fp}$ = 1 − specificity rate (% of real 0s predicted as 1)
- false negative rate $= \frac{fn}{fn+tp}$ (% of real 1s predicted as 0)

} depends on cut-off value, t or f(X).

(2) order observations based on predictions descendingly, i.e. find $k_1, k_2 \dots k_n$ s.t. $f(X_{k_1}) \geq f(X_{k_2}) \dots \geq f(X_{k_n})$ where $f(X)$ represents the prob. of observation being classified as 1. e.g.

$f(X_i)$	Y_i	R_s (t=.5)	R_{fp} (t=.5)
0.53	1	1/50	0/50
0.52	1	2/50	0/50
0.51	1		
⋮	⋮	⋮	⋮
0.5	1	25/50	0/50
0.49	0	25/50	1/50
0.48	1	26/50	1/50
0.47	0		2/50
⋮	⋮	⋮	⋮
0.01	0	50/50	32/50

n=100, n_p = 50, n_n = 50

(3) compute sensitivity rate $R_{s,j} = \frac{tp_j}{tp_j + fn_j} = \frac{\sum_{i=1}^{j} Y_{k_i}}{\sum_{k=1}^{n_p} Y_k}$ and

false positive rate $R_{fp,j} = \frac{fp_j}{fp_j + tn_j} = \frac{\sum_{i=1}^{j} (1-Y_{k_i})}{\sum_{k=1}^{n_n} (1-Y_k)}$, $j=1,\dots n$.

(4) plot $R_{s,j}$ against $R_{fp,j}$, $j=0,\dots n$

(5) let t be the cut-off value, the point on ROC $(R_s(t), R_{fp}(t))$ is:

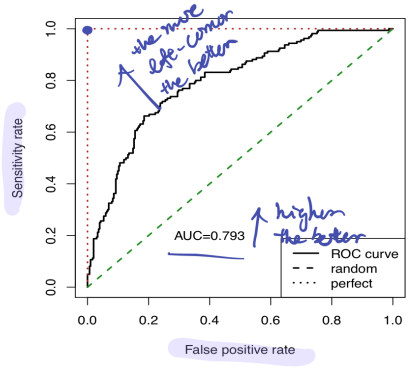
$$R_s(t) = \frac{\sum_{i=1}^{n} Y_i I(f(x_i) > t)}{\sum_{i=1}^{n} Y_i}, \quad R_{fp}(t) = \frac{\sum_{i=1}^{n} (1-Y_i) I(f(x_i) > t)}{\sum_{i=1}^{n} (1-Y_i)}$$

(6) AUC $= \frac{1}{2} \sum_{j=0}^{n-1} (R_{s,j} + R_{s,j+1})(R_{fp,j+1} - R_{fp,j})$.

compare with 'perfect curve' and 'random curve'
- perfect curve: prediction is observation itself i.e. $fp = 0$, $tp = 1$
- random curve: randomly prediction i.e. 45° line

relationship with Gains curve:
- Gains curve is a 'flattened' version of ROC.
- if one model dominates on one, it dominates on the other as well.

problem: ignore the actual probability but focus on the ordering.



(3) AvE

steps: (1) categorise X_{ij} into X_{ij}^*, j: X_{ij}^* in k^{th} category, $i = 1, \dots n$

(2) $\text{average}_k = \frac{\sum_{i=1}^n Y_i I(X_{ij}^* = k)}{\sum_{i=1}^n I(X_{ij}^* = k)}$, $\text{expected}_k = \frac{\sum_{i=1}^n \hat{f}(X_i) I(X_{ij}^* = k)}{\sum_{i=1}^n I(X_{ij}^* = k)}$.

(3) plot average_k against expected_k. $k = 1, \dots k$.

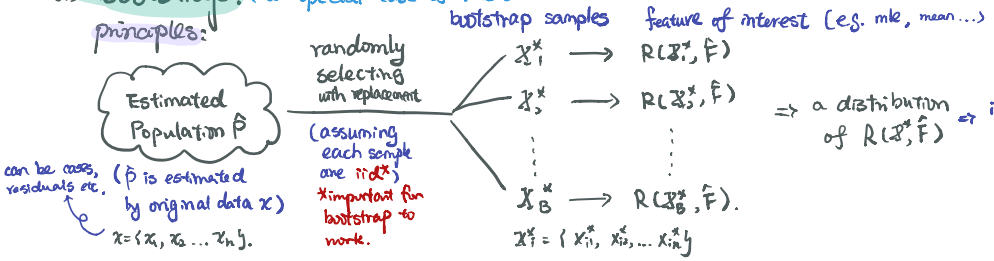
(4) replace X_{ij}^* with $f(X_{ij}^*)$ to plot AvE for $f(X)$

(4) Partial dependence plots

methods: Suppose we have p predictors, split them into X_S and X_C. So, $f(X) = f(X_S, X_C)$
Then the partial dependence of Y on X_S is defined as $f_S(X_S) = E_{X_C} f(X_S, X_C)$ with empirical approximation $\tilde{f}_S(X_S) = \frac{1}{n}\sum_{i=1}^n f(X_S, X_{i,C})$
Then we can plot $\tilde{f}_S$ against X_S.

(1) Bootstrap. (a special case of MC)

principles:



note:
- bootstrap cannot estimate prediction error because about 2/3 of original data appear in each bootstrap sample.
  => bootstrap will seriously underestimate prediction error as the model is trained with 2/3 of data.
- there total n^n possible samples, in which $\binom{2n-1}{n}$ are different. <=> $B << \binom{2n-1}{n}$.

Theoretical Background:
- True population $X \sim F$, we can use the ecdf of the original sample $\hat{F}$ to estimate F.
- Thus $R(X^*, \hat{F})$ can be an estimator of $R(X, F)$, which is a plug-in estimator, e.g. $\frac{1}{n}\sum_{i=1}^n X_i$ is a plug-in estimator of population mean, $\frac{1}{n}\sum_{i=1}^n X_i^2 - (\frac{1}{n}\sum_{i=1}^n X_i)^2$ is a plug-in estimator of population variance.
- The bootstrap principle says $R(X, F) \overset{d}{\approx} R(X^*, \hat{F})$.

Methods:

(1) Nonparametric & parametric:

Nonparametric: (1) Bootstrap samples are drawn from $x = \{x_1, \dots x_n\}$ directly => $x_i^* = \{x_{i1}^*, \dots x_{in}^*\}$ for $i = 1, \dots B$
(2) Then ecdf of $\{R(X_i^*, \hat{F}), i = 1, \dots B\}$ is used to estimate cdf of $R(X^*, \hat{F})$ and thus $R(X, F)$.

Parametric: (1) Assume $X_1 \dots X_n \overset{iid}{\sim} F(x|\theta)$, which can be estimated by $\hat{\theta}$ instead of being estimated by ecdf $\hat{F}$.
<=> (e.g. mle, mom...)
Then draw bootstrap samples from $F(x|\hat{\theta})$ assuming each sample is iid. => $x_i^* = \{x_{i1}^* \dots x_{in}^*\} \overset{iid}{\sim} F(x|\hat{\theta})$
(2) Then ecdf of $\{R(X_i^*, F(x|\hat{\theta}))\}$ is used to estimate cdf of $R(X^*, \hat{F}(x|\hat{\theta}))$ and thus $R(x, F(x|\theta))$

Comparison between nonparametric & parametric:

Approx. of	Nonparametric BS	Parametric BS
# samples	B samples randomly drawn from $\mathcal{X}$	
true dist.	$\mathcal{X} = \{x_1, \dots x_n\} \overset{iid}{\sim} F \approx \hat{F}$	$\mathcal{X} = \{x_1, \dots x_n\} \overset{iid}{\sim} F(x\mid\theta) \approx F(x\mid\hat{\theta})$
BS dist.	$\mathcal{X}_i^* = \{x_{i1}^*, x_{i2}^*, \dots x_{in}^*\} \overset{iid}{\sim} \hat{F}$ for $i=1,\dots B$	$\mathcal{X}_i^* = \{x_{i1}^*, x_{i2}^*, \dots x_{in}^*\} \overset{iid}{\sim} F(x\mid\hat{\theta})$ for $i=1,\dots B$
ideal BS dist.	$R(\mathcal{X}^*, \hat{F}) \approx$ ECDF of $\{R(\mathcal{X}_i^*, \hat{F}), i=1,\dots B\} = \hat{F}^*$	$R(\mathcal{X}^*, F(x\mid\hat{\theta})) \approx$ ECDF of $\{R(\mathcal{X}_i^*, F(x\mid\hat{\theta})), i=1,\dots B\} = \hat{F}^*$
quantity dist.	$R(\mathcal{X}, F) \approx \hat{F}^*$	$R(\mathcal{X}, F(x\mid\theta)) \approx \hat{F}^*$

(2) Bootstrap in regression:

context: consider a regression model: $Y_i = x_i^T\beta + \varepsilon_i$ for $i=1,\dots n$ and $\varepsilon_i \overset{iid}{\sim} F(0,\sigma^2)$.

observed data are $\{z_1 = (x_1, y_1) \dots z_n (x_n, y_n)\}$

two ways to bootstrap sample from observed data are: bootstrap the residuals and bootstrap the case.

2.1 bootstrap the residuals (nonparametric/parametric).

steps:

all these must be done in one process:

(1) fit the model to the observed data. Obtain $\hat{y}_i = x_i^T\hat{\beta}$ and $\hat{\varepsilon}_i = y_i - \hat{y}_i$

(2) nonparametric: bootstrap the residuals $\{\hat{\varepsilon}_1, \dots \hat{\varepsilon}_n\}$ to obtain $\{\hat{\varepsilon}_1^*, \dots \hat{\varepsilon}_n^*\}$

(2)* parametric: simulate $\{\hat{\varepsilon}_1^*, \dots \hat{\varepsilon}_n^*\}$ from some distribution that we assume } (note: important to check if $\{\hat{\varepsilon}_1, \dots \hat{\varepsilon}_n\}$ are iid.)

(3) create a bootstrap response: $Y_i^* = \hat{y}_i + \hat{\varepsilon}_i^*$ for $i=1,\dots n$

(4) fit the model to $\{(x_i, Y_i^*), i=1,\dots B\}$ to get bootstrap estimates of $\hat{\theta}^*$

(5) repeat this for B times to get $\{\hat{\theta}_1^*, \dots \hat{\theta}_B^*\}$, from which an ecdf can be built for statistical inference.

2.2 bootstrap the cases (nonparametric).

steps:

all these must be done in one process:

(1) bootstrap the sample $\{z_1, \dots z_n\}$ to obtain $\{z_1^* \dots z_n^*\}$ (note: here we assume $\{z_1, \dots z_n\}$ are iid).

(2) fit the model to $\{z_1^*, \dots z_n^*\}$ to get bootstrap estimates of $\hat{\theta}^*$.

(3) repeat this for B times to get $\{\hat{\theta}_1^*, \dots \hat{\theta}_B^*\}$, from which an ecdf can be built for statistical inference.

* Bootstrapping the cases is less sensitive to violations in the regression model assumptions. i.e. adequacy of the model and constancy of σ^2) than bootstrapping the residuals.

2.3 bootstrap estimation of bias($\hat{\theta}$) and se($\hat{\theta}$).

$$\text{bias}(\hat{\theta}) = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^* - \hat{\theta} = \bar{\theta}^* - \hat{\theta} \qquad \text{se}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i^* - \bar{\theta}^*)^2}$$

(3) Bootstrap inference contents: (Bootstrap CI).

3.1 Normal

$$\left[2\hat{\theta} - \bar{\hat{\theta}}^* - z_{1-\alpha/2}\cdot\text{se}(\hat{\theta}),\ 2\hat{\theta} - \bar{\hat{\theta}}^* + z_{1-\alpha/2}\cdot\text{se}(\hat{\theta})\right]$$

$2\hat{\theta} - \bar{\hat{\theta}}^* = \hat{\theta} - \text{bias}(\hat{\theta}) = \hat{\theta} - (\bar{\hat{\theta}}^* - \hat{\theta})$

$= 2\hat{\theta} - \bar{\hat{\theta}}^*$ = unbiased estimate of θ.

✓ follow the 'generic template'

✗ may be used for large sample size or normal population.

3.2 Percentile (or basic percentile)

$$\left[\hat{\theta}^*_{\alpha/2(B+1)},\ \hat{\theta}^*_{1-\alpha/2(B+1)}\right]$$

✗ highly asymmetric

✗ vulnerable to bias & inaccurate coverage prob.

✓ may work for location parameters.

3.3 Basic (or residual)

$$\left[2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2(B+1)},\ 2\hat{\theta} - \hat{\theta}^*_{\alpha/2(B+1)}\right]$$

✓ adjusted for bias.

3.4 BCa (Bias corrected and accelerated).

$$\left[\hat{\theta}^*_{(B+1)P_L},\ \hat{\theta}^*_{(B+1)P_U}\right]$$

where $P_L = \Phi\left(\frac{C_0 - z_p}{1-a(C_0-z_p)} + C_0\right)$, $P_U = \Phi\left(\frac{C_0+z_p}{1-a(C_0+z_p)} + C_0\right)$,

C_0 is determined by the relative position of $\hat{\theta}$ among $\{\hat{\theta}^*_1, \dots \hat{\theta}^*_B\}$

a is determined by skewness of $\{\hat{\theta}^*_1, \dots \hat{\theta}^*_B\}$

✓ adjusted for bias & skewness ✓ reasonable computation time. × need EIF for parametric BS.

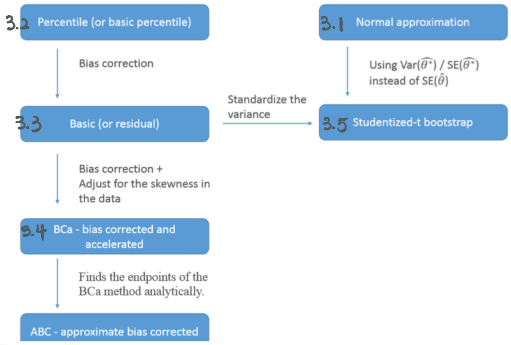
3.5 Studentized

$$\left[\hat{\theta} - \xi_{1-\alpha/2}(\hat{G}^*)\sqrt{\hat{V}(\hat{\theta})},\ \hat{\theta} - \xi_{\alpha/2}(\hat{G}^*)\sqrt{\hat{V}(\hat{\theta})}\right]$$

where $\hat{G}^*$ is the distribution of $R(\mathcal{X}^*, \hat{\theta}) = \frac{\hat{\theta}^* - \hat{\theta}}{\sqrt{V(\hat{\theta}^*)}}$,

$\xi_\alpha(\hat{G}^*)$ is the α quantile of $\hat{G}^*$.

✓ symetrical CI by using $\frac{\hat{\theta}^* - \hat{\theta}}{\sqrt{V(\hat{\theta}^*)}}$.
✓ reasonable for location and variance-stabilized estimators.
× sensitive to presence of outliners
× performs badly for heavy tailed dist.
× computationally expensive - requires double bootstrap to calculate $\hat{V}(\hat{\theta}^*)$.



(4) Bootstrap hypothesis testing.

$H_0: \theta = \theta_0$ vs. $H_1: \theta \neq \theta_0$

can use statistics $\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\theta}^*}$ or can use permutation test.

reducing MC error: ($\frac{1}{B}\sum_{j=1}^{B}(\bar{X}^*_j - \bar{X})$ is unlikely to be 0 due to MC variation).

(1) balanced bootstrap: (all observations occur with the same frequency as in the original sample)

The simplest way to get B balanced bootstrap samples is to concatenate B copies of the observed sample of size n, randomly permute this series, and then read off B blocks of size n sequentially. The jth block becomes the jth bootstrap sample $\mathcal{X}^*_j$. For the permutation involved, balanced bootstrap is also called **permutation bootstrap**.

(2) antithetic bootstrap:

- For a sample of univariate data, $x_1, \cdots, x_n$, denote the ordered data as $x_{(1)}, \cdots, x_{(n)}$. Let $\pi(i) = n - i + 1$ be a permutation operator that reverses the order statistics.
- Then for each bootstrap sample $\mathcal{X}^* = \{X^*_1, \cdots, X^*_n\}$, let $\mathcal{X}^{**} = \{X^{**}_1, \cdots, X^{**}_n\}$ denote the sample obtained by substituting $X_{(\pi(i))}$ for every instance of $X_{(i)}$ in $\mathcal{X}^*$. Thus, for example, if $\mathcal{X}^*$ has an unrepresentative predominance of the larger observed data values, the smaller observed values will predominate $\mathcal{X}^{**}$.
- Using this strategy, each bootstrap draw provides two estimators: $R(\mathcal{X}^*, \hat{F})$ and $R(\mathcal{X}^{**}, \hat{F})$. The two estimators are often negatively correlated.

e.g. $B=3$, $n=3$

$\mathcal{X}^*$: 112 223 133 → $R(\mathcal{X}^*)$

$\mathcal{X}^{**}$: 332 221 311 → $R(\mathcal{X}^{**})$

permutation BS samples: $R(\mathcal{X}^*)$

antithetic BS samples: $R_a(\mathcal{X}^*) = \frac{1}{2}[R(\mathcal{X}^*) + R(\mathcal{X}^{**})]$

- Let $R_a(\mathcal{X}^*, \hat{F}) = \frac{1}{2}[R(\mathcal{X}^*, \hat{F}) + R(\mathcal{X}^{**}, \hat{F})]$. Then R_a has the following desirable property

$$\begin{aligned}\mathrm{Var}[R_a(\mathcal{X}^*, \hat{F})] &= \frac{1}{4}\Big([\mathrm{Var}[R(\mathcal{X}^*, \hat{F})] + \mathrm{Var}[R(\mathcal{X}^{**}, \hat{F})] \\ &\quad + 2\mathrm{Cov}[R(\mathcal{X}^*, \hat{F}), R(\mathcal{X}^{**}, \hat{F})]\Big) \\ &\leq \mathrm{Var}[R(\mathcal{X}^*, \hat{F})]\end{aligned}$$

if the covariance is negative.

- The above strategy of reducing Monte Carlo error in bootstrap is referred to as **antithetic bootstrap**.
- It is also possible to establish ordering of multivariate data to permit an antithetic bootstrap strategy.