Валидиране на модели

Статистически характеристики

Кръстосано валидиране: LOO, LMO

Boot-strapping

Y-scrabling

Разделяне на обучителната извадка

По същество всички модели са грешни, но някои модели са полезни!

George E. P. Box

Колко “грешен” трябва де е моделът за да НЕ е полезен?

- Всички модели са непълни, но в някои има полезни идеи.
- Всички модели са приближения, но някои са полезни.

Описание на модела и извадките с данни

Недвусмислени методи за моделиране:

ULR - линейна регресия с една променлива
MLR - линейна регресия с много променливи
PCA - анализ на главните компоненти
PCR - регресия по главаните компоненти
PLS - метод на частичните най-малки квадрати
ANN - изкуствени невронни мрежи
"Размито" кластериране и регресия (Fuzzy Clustering)
KNN - метод на най-близките K съседа
GA - генетични алгоритми

в общ вид моделът може да се опише:

$$Y = F(X)$$

Описание на модела и извадките с данни

Структури

Целево свойство за моделиране

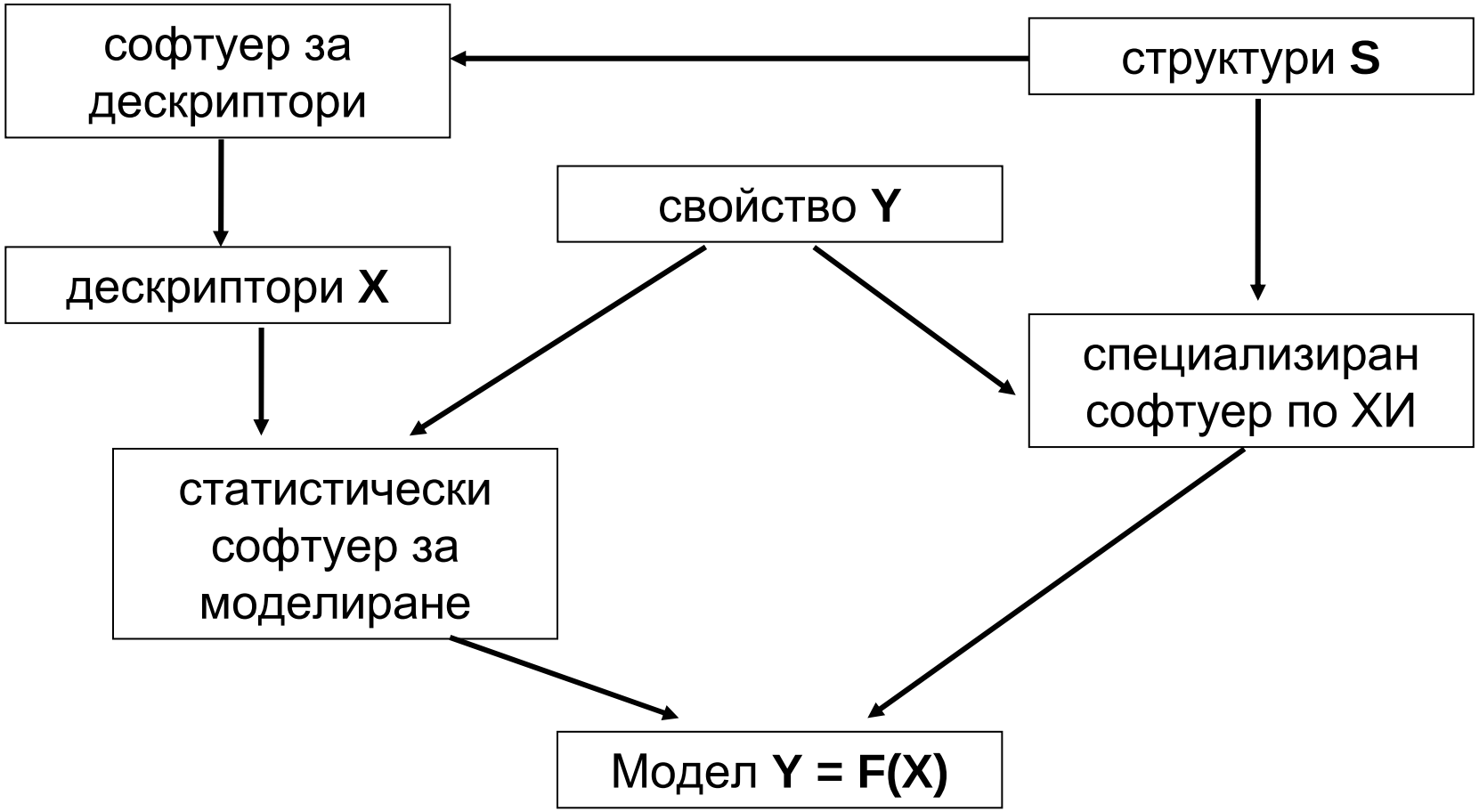
Матрица с дескрипторите

$$S = \begin{pmatrix} S_1 \\ S_2 \\ . \\ . \\ S_n \end{pmatrix} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} x_{1,1} & x_{2,1} & x_{3,1} & \cdots & x_{m,1} \\ x_{1,2} & x_{2,2} & x_{3,2} & \cdots & x_{m,2} \\ & & & & \\ & & & & \\ x_{1,n} & x_{2,n} & x_{3,n} & \cdots & x_{m,n} \end{pmatrix}$$

Всяка извадка с данни се състои от следните компоненти: **S, Y, X**

Описание на модела и извадките с данни

Дескрипторите **X** в общия случай са функция от структурата на съединението **X = f(S)**



Работни извадки с данни

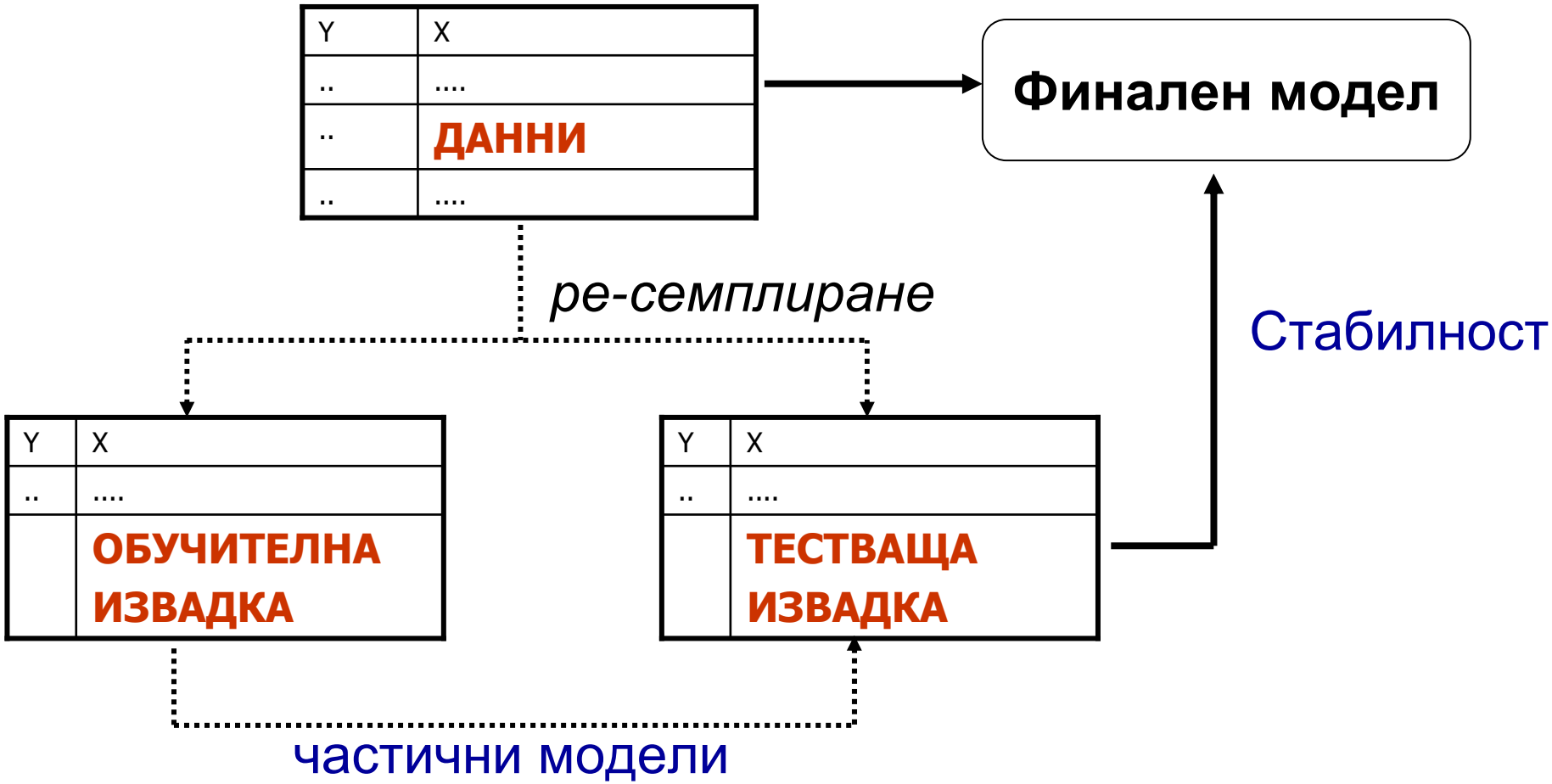
При създаване на модели се работи с 3 вида извадки от химични обекти:

- обучителна извадка - данните се използват за определяне на модела (получаване на знание)

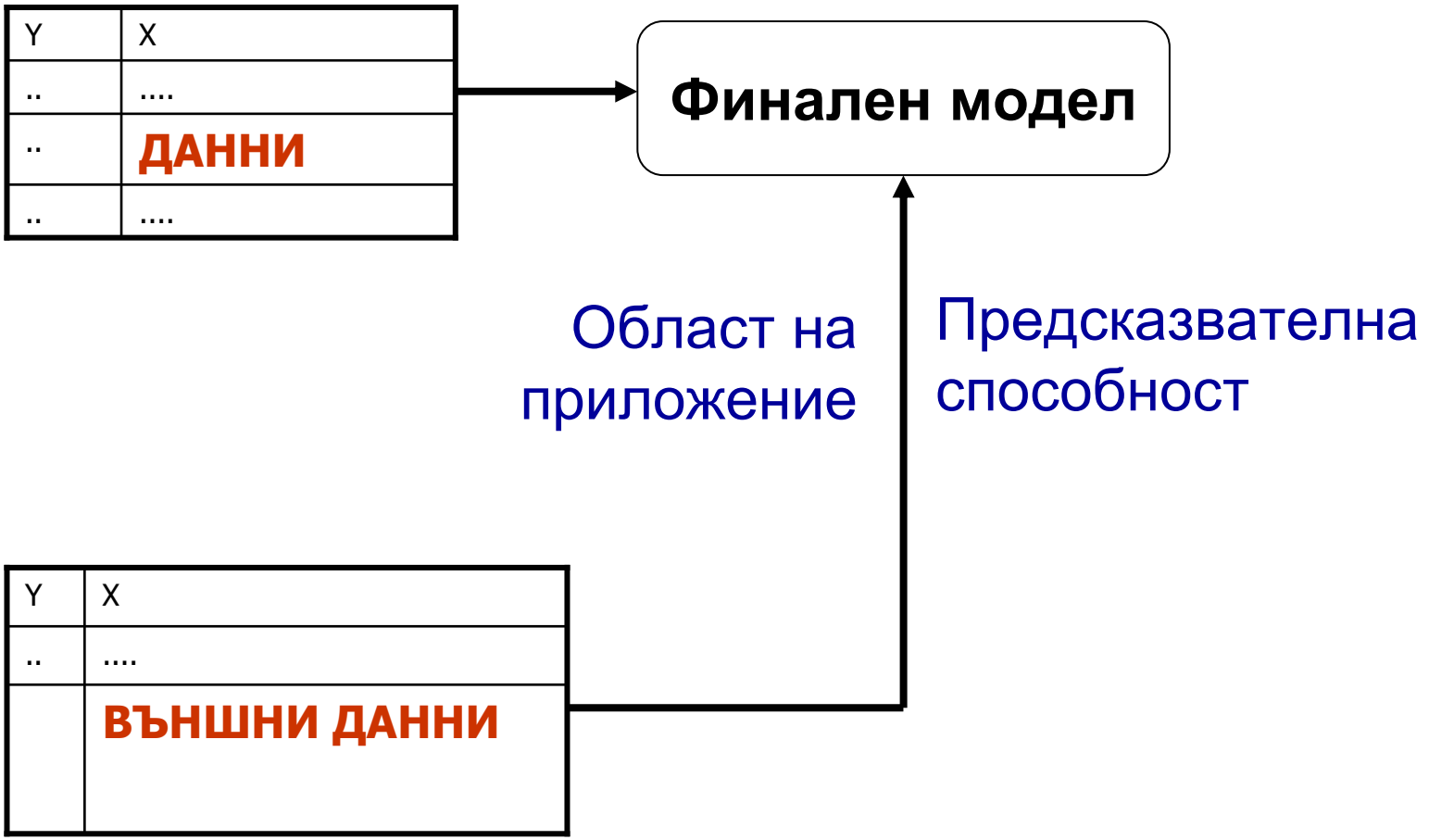
- валидираща извадка - данните се използват за статистическо охарактеризиране/валидиране на модела

- тестваща извадка - данните се използват за допълнителни тестове. Тези данни са от друг външен източник

Вътрешно валидиране на модел



Външно валидиране на модел



Валидиране на модели

При валидирането на модел се доказва:

- качество на приближението (goodness-of-fit)

- статистическа значимост на изчислените параметри

- стабилност (robustness)

- предсказвателна способност (predictability)

- преносимост (област на приложение, Applicability Domain)

Характеристики на модел

Обозначения: $\hat{y} = b_0 + b_1 x_1 + \ldots + b_m x_m$

$\hat{y}_i$ моделирана (изчислена) стойност

y_i експериментална стойност

x_i дескриптор (предсказваща променлива)

b_i коефициент в регресията

$\bar{y}$ средно на експерименталните стойности $\{y_i\}$

$\hat{\bar{y}}$ средно на моделираните стройности

Характеристики на модел

Основните характеристики на даден модел се получават след сравняване на вектор-колоните с експерименталните и моделираните стойности

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \qquad \hat{Y} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ . \\ . \\ \hat{y}_n \end{pmatrix}$$

Помощни суми

Пълна сума от квадратите (total sum of squares):

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Сума от остатъците на квадрат (residual sum of squares):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Сума на квадратите дължащи се на регресията (explained sum of squares)

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Помощни суми

Връзка между сумите с квадратите:

$$TSS = ESS + RSS$$

“Тоталната” (пълната вариация) се определя като сума на “обяснената” вариация дължаща се на регресионния модел и сумата от грешките (остатъците) - необяснената вариация.

Класически характеристики на модел

Средно-квадратична грешка:

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

Класически характеристики на модел

Класически коефициент на корелация $r = \frac{\mathrm{cov}(\hat{Y}, Y)}{S_{\hat{Y}}.S_Y}$

$$r = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson Product-Moment Correlation Coefficient (PMCC)

Класически коефициент на корелация r^2

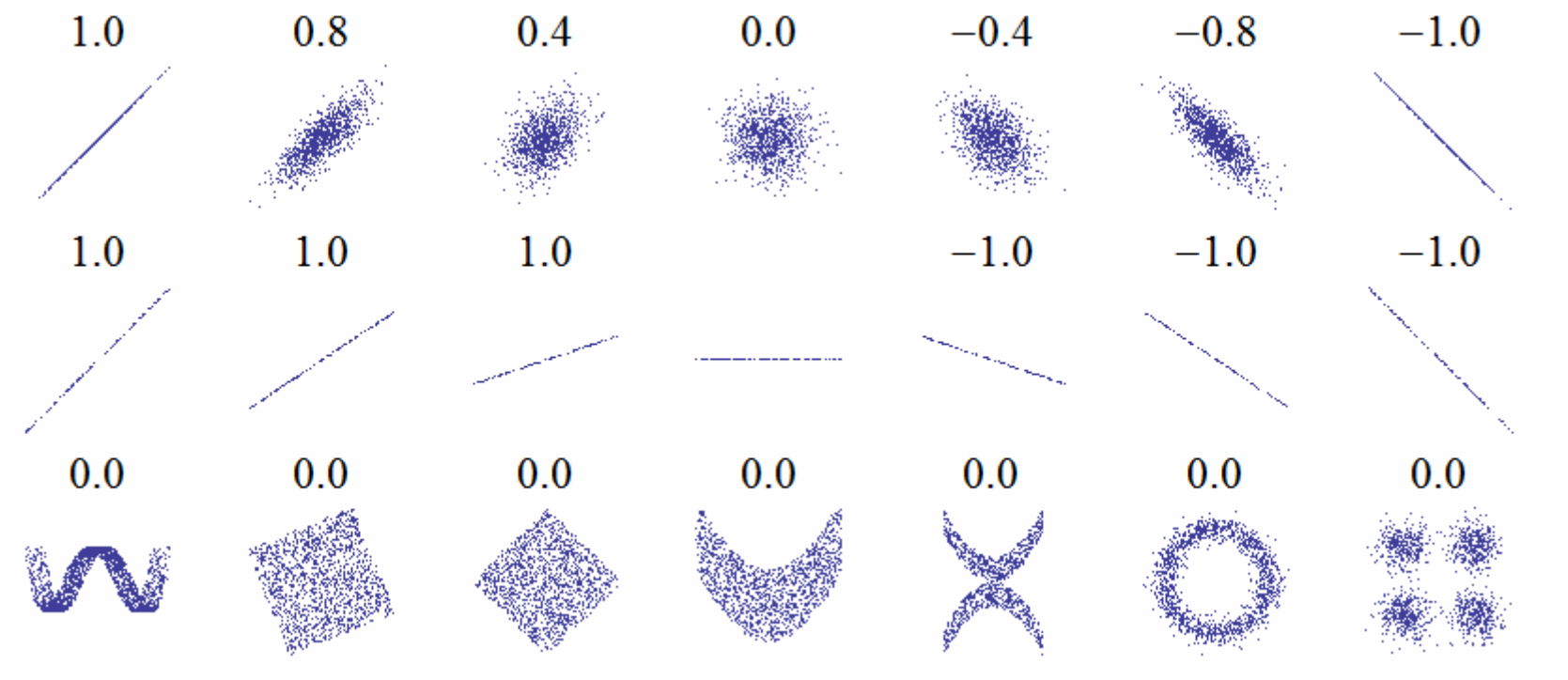
В официалните изисквания за QSAR и QSPR моделиране **не се препоръчва** употребата на класическия коефициент на корелация $\mathbf{r^2}$.

Използват се **други дефиниции** за коефициент на корелация при валидирането и тестването на QSAR и QSPR модели.

При много точни модели (например при калибрация) класическият коефициент на корелация и алтернативните коефициенти на корелация практически съвпадат.

Класически коефициент на корелация r^2

Основно предназначение на класическия коефициент на корелация **r^2** е определянето степента на линейна зависимост между две променливи.



Класически коефициент на корелация r^2

Недостатъци при валидиране на модел:

При тестване с външна извадка, във формулата за $\mathbf{r^2}$ участват статистически параметри свързани със разпределението на външната тестваща извадка.

Коефициентът на корелация $\mathbf{r^2}$ зависи от начина, по който е конструирана тестващата извадка, следователно не е подходящ за определяне на предсказвателната способност на модели.

Класически коефициент на корелация r^2

Недостатъци:

Много добре известен дефект на класическия коефициент на корелация е следният факт:

ако моделът е точна линейна функция от експерименталните стойности:

$$\hat{y}_i = ay_i + b$$

то

$$r^2 = 1$$

независимо от големината на отклоненията между моделираните и експерименталните стойност.

Класически коефициент на корелация r^2

Коефициентът **r^2** може да се използва за оценка

не на качеството на моделираните стойности,

а за оценка на възможността да се състави модел от избраните дескриптори.

Класически коефициент на корелация r^2

Внимание!

Корелацията не означава причинно-следствена връзка.

Характеристики на модел

Алтернативен коефициент на корелация - **коефициент на определяне** (coefficient of multiple determination):

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Коефициентът на определяне $\mathbf{R^2}$ е по-критичен от класическият коефициент на корелация $\mathbf{r^2}$. При по-слаби корелации, $\mathbf{R^2}$ дава по-ниски (по-песимистични стойности).

Характеристики на модел

Коефициентът на определяне

$$R^2 = 1 - RSS/TSS$$

може да се интерпретира като процента на вариацията, която е обяснена от дескрипторите на модела.

Процент на необяснената вариация

$$FVU = 1 - R^2 = RSS/TSS$$

Характеристики на модел

При валидиране и тестване с външни извадки експерименталната стойност y_i не участва в модела, чрез който се получава $\hat{y}_i$

В тези случай сумата **RSS** се обозначава като **PRESS** (Predicted Residual Sum of Squares).

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Характеристики на модел

При валидиране и тестване с външни извадки се изчислява **коефициент на предсказване**:

$$Q^2 = 1 - \frac{PRESS}{TSS}$$

Характеристики на модел

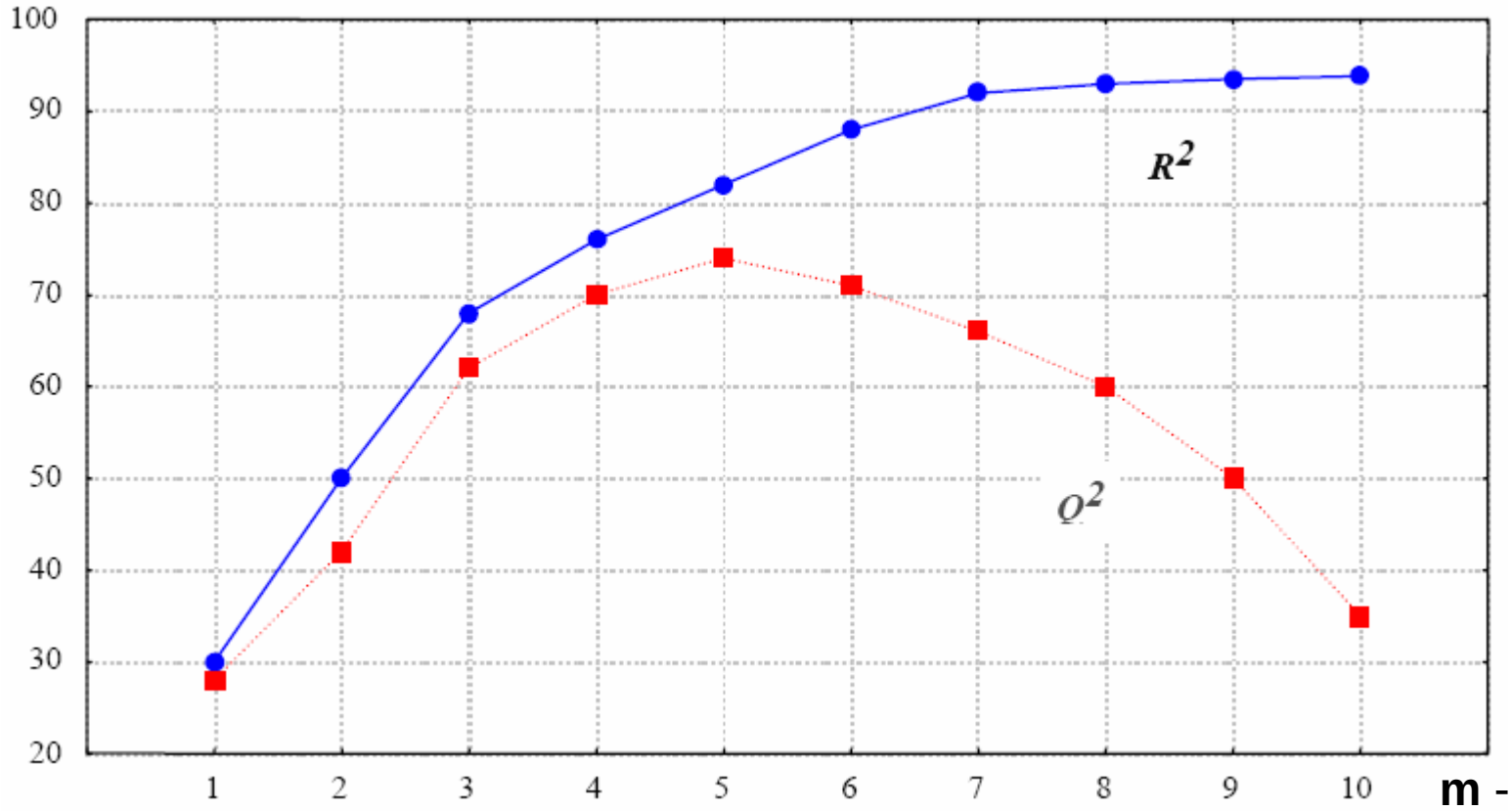
Коефициентът на определяне **R^2** характеризира качеството на приближението на модела (goodness-of-fit).

Коефициентът на предсказване **Q^2** определя предсказвателната способност на модела (predictability).

$$Q^2 \leq R^2 \leq 1$$

Получаването на високи стойности на **Q^2** е подсигуряване против ефекта на “премоделирането”.

Сравняване на R^2 и Q^2



m - брой променливи в модела

Премоделиране (over-fitting)

Статистическият модел е премоделиран, когато се описва случайната грешка и шума вместо истинските взаимовръзки в данните.

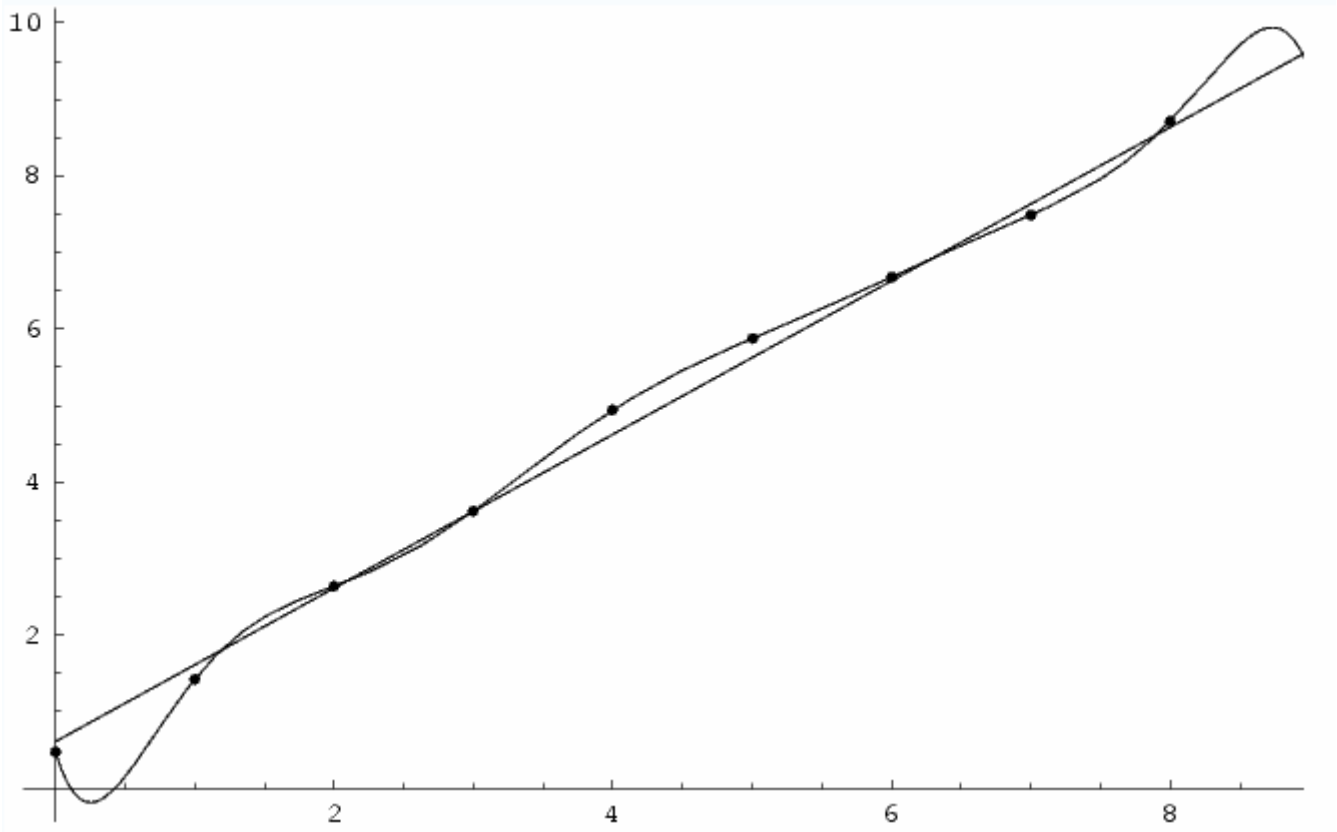
Премоделиране се получава когато моделът е много сложен - с голям брой степени на свобода спрямо количеството данни в обучителната извадка.

Премоделираният модел има ниски стойности за $\mathbf{Q^2}$

Премоделиране може да се получи не само от голям брой степени на свобода, но и от цялостната концепция на модела, ако се среми да се намали грешката на модела дължаща се на случайния шум.

Премоделиране

Премоделиране дължащо се на апроксимиране на даните с полином вместо с линейна функция



Премоделиране

Премоделирането се избягва чрез различни методи за валидиране!!!

Характеристики на модел

"Коригиран" коефициент на определяне:

$$R^2_{adj} = 1 - \frac{RSS/(n-m-1)}{TSS/(n-1)}$$

$$R^2_{adj} = 1 - (1-R^2)\frac{(n-1)}{n-m-1)}$$

Характеристики на модел

Коригираният коефициент на определяне R^2_{adj} се интерпретира по различен начин от основната дефиниция $R^2 = 1 - RSS/TSS$

Коригираният коефициент R^2_{adj} нараства, само ако новите добавени дескриптори статистически значимо подобряват модела.

Коригираният коефициент R^2_{adj} играе ролята на наказателна функция при избиране на неподходяши дескриптори, водещи до премоделиране.

Характеристики на модел

Стандартна грешка на модела: $$s = \sqrt{\frac{RSS}{n-m-1}}$$

F-тест $$F = \frac{ESS / m}{RSS /(n-m-1)}$$

Характеристики на модел

Интервална оценка на коефициентите на регресията:

t-тест

$$t = \frac{b_i}{s(b_i)}$$

стандартно отклонение на коефициента $\boldsymbol{b_i}$

$$s(b_i) = \sqrt{\frac{RSS\ /(n-m-1)}{\sum_{k=1}^{n}(x_{i,k} - \overline{x}_i)^2}}$$

Характеристики на модел

При тестване с външна извадка са дефинирани две основни разновидности на коефициента на предсказване:

$$Q^2_{F1} = 1 - \frac{PRESS}{SS_{EXT}(\bar{y}_{TR})} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2}$$

$$Q^2_{F2} = 1 - \frac{PRESS}{SS_{EXT}(\bar{y}_{EXT})} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2}$$

Характеристики на модел

В официалната документация на OECD за QSAR моделиране се препоръчва употребата на Q^2_{F1}

Оценяването на външен тест чрез коефициента Q^2_{F1} е адекватно само когато експерименталноите стойности на тестващата извадка са разпределни както стойностите на обучаващата извадка.

Характеристики на модел

Недостатъци на коефициента Q^2_{F1}

Предсказвателната способност на модела се надценява, когато тестовите обекти са с експериментални стойности в граничните региони на разпределението на стойнотите от обучителната извадка.

Това поведение на Q^2_{F1} е обратно на логиката на областта на приложимост на даден модел.

В някои случаи Q^2_{F1} има стойности по-големи от R^2

Характеристики на модел

Недостатъците на Q^2_{F1} са подобрени в дефиницията на Q^2_{F2} .

От свойствата на сумите SS_{EXT} следва зависимостта:

$$Q^2_{F2} \leq Q^2_{F1}$$

Като цяло Q^2_{F2} също има недостатъци и зависи от начина, по който е конструирана тестващата извадка.

Характеристики на модел

Групата на проф. Тодесчини предлага вариант на коефициента **Q^2**, който оценява предсказващата способност на модела без да се влияе от конструирането на външната тестваща извадка

$$Q^2_{F3} = 1 - \frac{PRESS / n_{EXT}}{TSS / n_{TR}} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n_{TR}} (y_i^{TR} - \bar{y}_{TR})^2}$$

Самовалидиране на модел

В този случай се работи само с обучителната извадка. Няма специално отделени обекти за валидиране (тестване)

Обектите от обучителната извадка се използват и за валидиране. От тях се изчисляват **R^2** и **RMSE**

Такъв вид валидиране, има по-малка статистическа тежест. Препоръчително е числовите характеристики да се използват само като ориентировъчни резултати.

При самовалидирането, има опасност при неправилен избор на голям брой дескриптори да се получи ‘премоделиране’

Валидиране на модели

При наличие на голям брой обекти, извадката се разделя случайно на две подмножества – обучителна извадка, с която се изчислява моделът и валидираща извадка, с която се определят статистическите характеристики на модела.

За получаване и доказване на по-добра преносимост може да се използа значително по-голяма валидираща извадка, при която се запава случйният избор на обектите. Обучителната извадка от друга страна може да е по-малка по размер, но да е съставена целенасочено, така че да се отчетат основните тенденции в модела.

Кръстосано валидиране

При по-малък брой на обекти с известни експериметални свойства се прилага методът кръстосано валидиране (cross-validation)

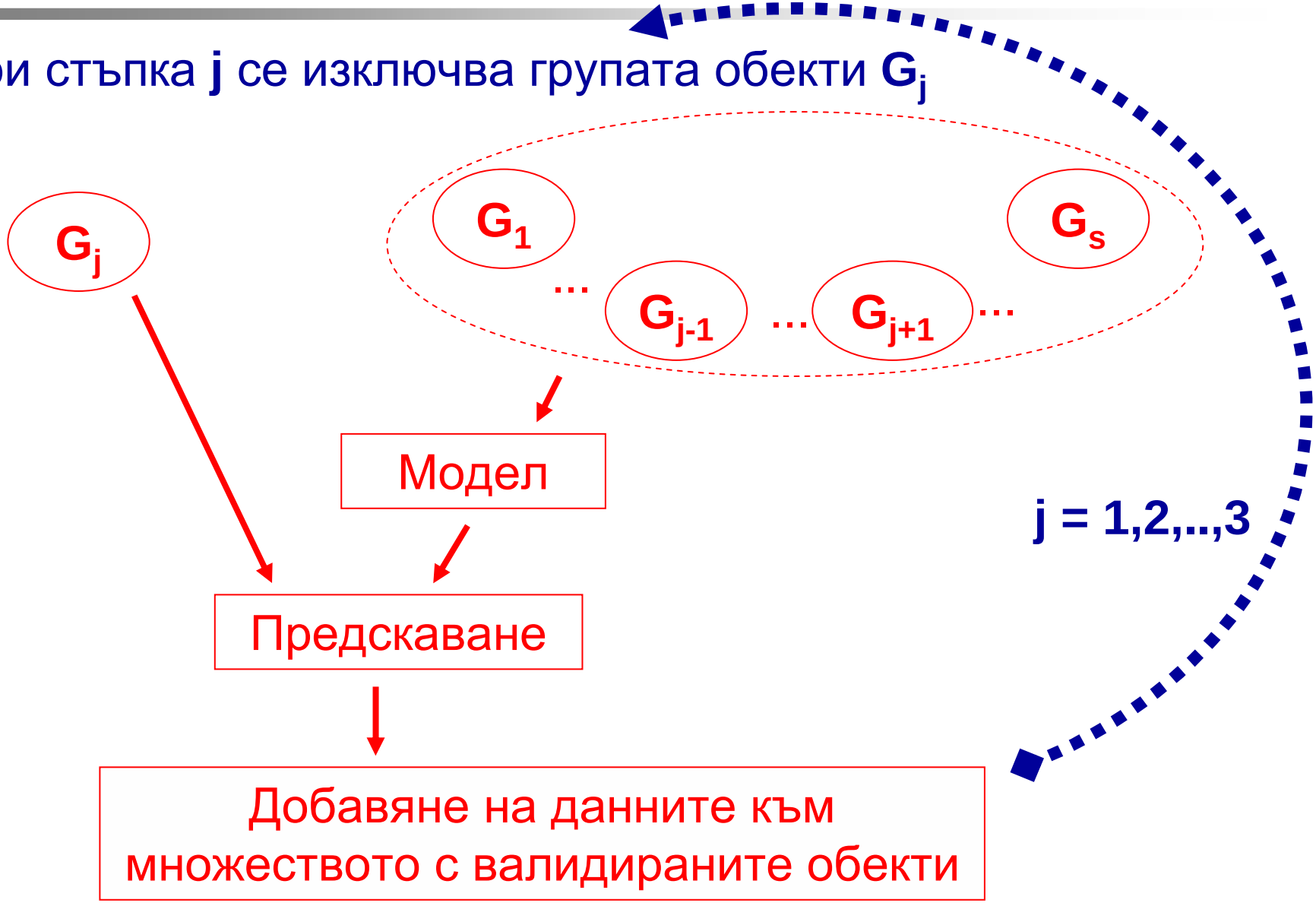
Наличната извадка с обекти се разделя на **s** групи от по **k** на брой обекти

Правят се **s** на брой итерации – при всяка итерация една от групите се изключва и се използва като валидираща извадка. С останалите обекти се прави модел, който се тества с изключените обекти.

Статистическата оценка на метода се прави като се обединят тестовите резултати от всичките **s** стъпки.

Кръстосано валидиране

При стъпка **j** се изключва групата обекти G_j



Кръстосано валидиране

При **k=1/4 n** методът се нарича leave-quarter-out (cross-validation)

При **k=1** методът се нарича leave-one-out (LOO cross-validation) или пълно кръстосано валидиране (full cross-validation)

При **k>1** методът се нарича leave-many-out (LMO cross-validation)

Bootstrap валидиране

При **bootstrap** методите се предполага, че обучителната извадка е достатъчно представителна за класа обекти които се изследват.

Обучителната извадка симулира генералната съвокупност от обкети т.е. когато се вземат обекти от оригиналната обучителната се приема че тя е генералната съвокупност

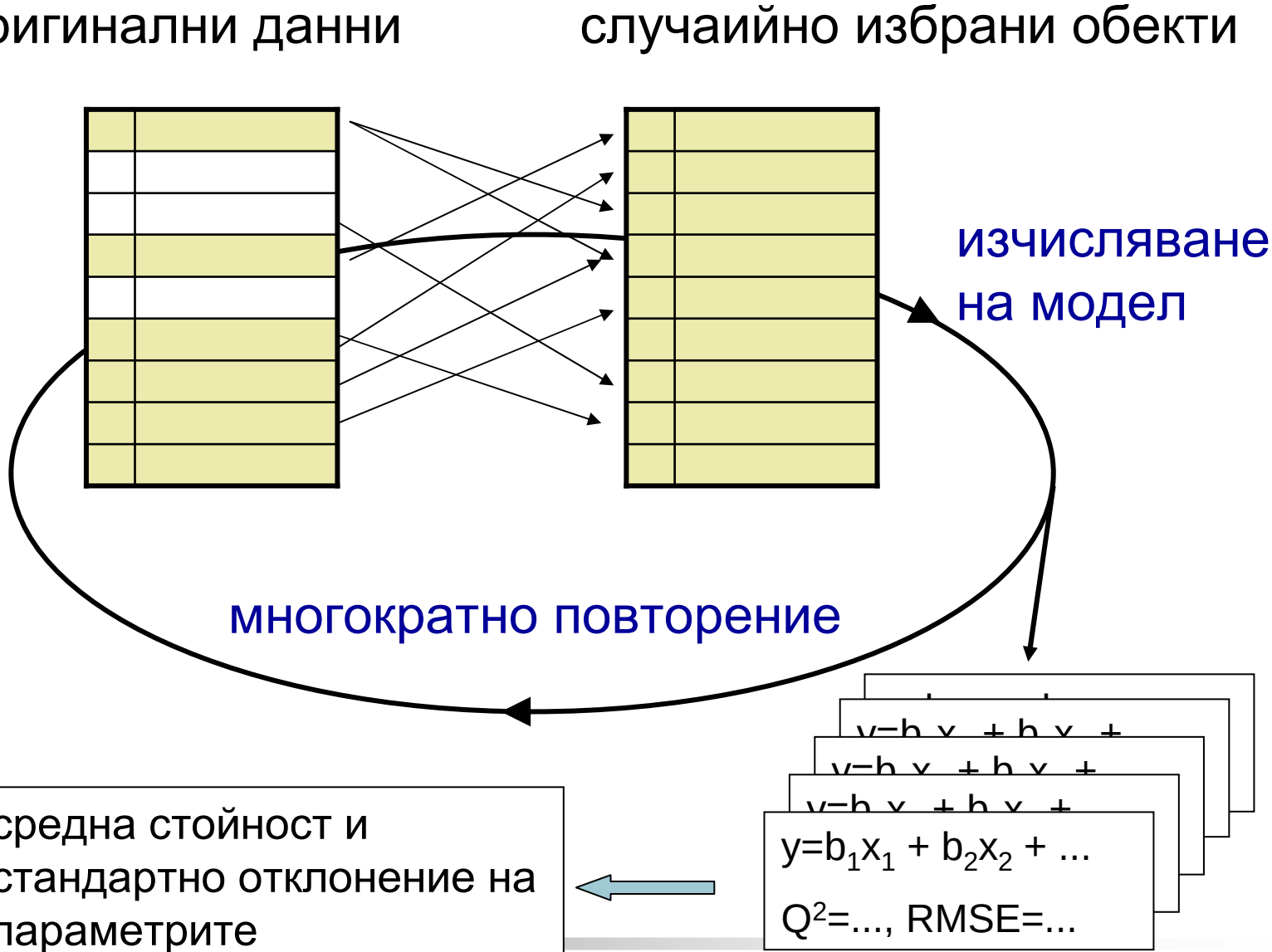
Bootstrap валидиране

Нови работни извадки наречени bootstrap извадки с размера на оригиналната зивадка се генерират, чрез случаен избор на обекти от оригиналната извадка.

Повторения на обектите в bootstrap извадките са позвололени.

При наличие на повторния в избраните обекти, част от оригиналните обекти не влизат в bootstrap извадката и именно те се използват за тестването на тази извадка.

Bootstrap валидиране



Bootstrap валидиране

Bootstrap тестовете се повтарят много пъти, например 5000 – 10000 повторения.

Средните стойности на статистическите параметри се използват за определяне на стабилността на модела.

$$Q^2_{BOOT} = \frac{1}{10000} \sum_{k=1}^{10000} Q^2{}_{(k)}$$

$$\overline{b_i} = \frac{1}{10000} \sum_{k=1}^{10000} b_i{}^{(k)}$$

Bootstrap валидиране

$\overline{b_i} - b_i$ мярка за систематичната грешка (bias) в оригиналния модел

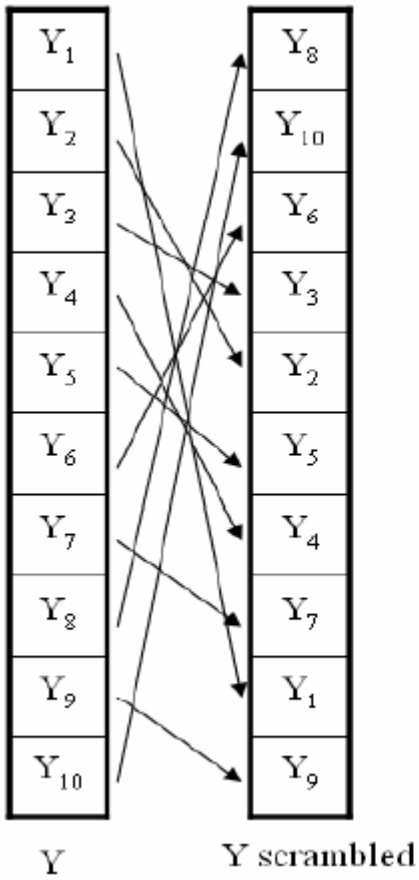
Y-scrambling валидиране

Y-scrambling (анг. буквално: y-разбъркване) е метод за излседване на стабилността на модела чрез произволна пермутация в обучителната извадка на стойностите на целевото свойство.

Процедурата се повтаря многократно като на всяка итерация се създава обучителна извадка, при която стройностите на дескрипторите (матрицата X) не се променят, а стойностите на вектор колоната Y са произволно разбъркани.

Y-scrambling валидиране

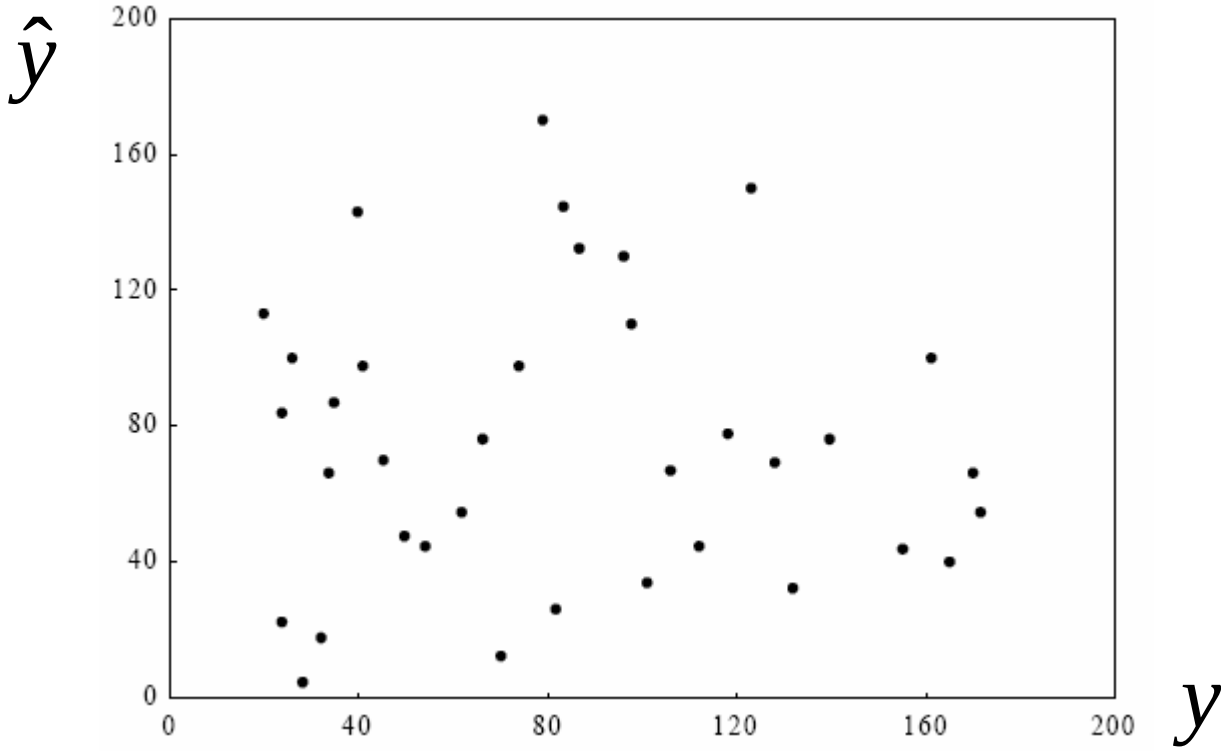
Съставяне на обучителна извадка при Y-scrambling



X_{11}	X_{21}	...	X_{n1}
X_{12}	X_{22}	...	X_{n2}
X_{13}	X_{23}	...	X_{n3}
X_{14}	X_{24}	...	X_{n4}
X_{15}	X_{25}	...	X_{n5}
X_{16}	X_{26}	...	X_{n6}
X_{17}	X_{27}	...	X_{n7}
X_{18}	X_{28}	...	X_{n8}
X_{19}	X_{29}	...	X_{n9}
X_{110}	$X_{2\ 10}$	...	X_{n10}

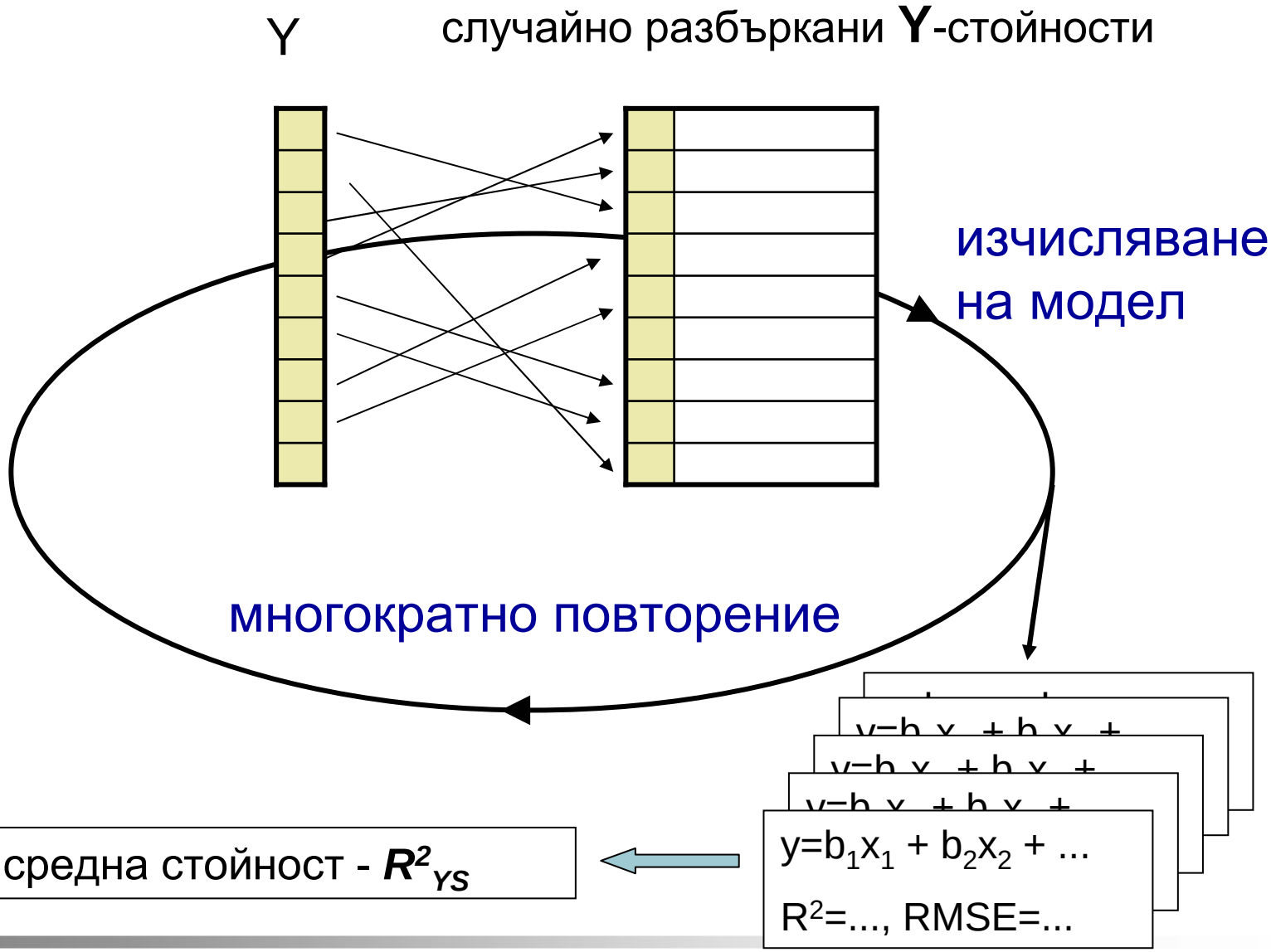
X

Y-scrambling валидиране



При сличайната пермутация се очаква стойностите на модела и експерименталните стройности да са произволно разбъркани в рванината т.е. $\mathbf{R^2}$ трябва да е много нисък.

Y-scrambling валидиране



Y-scrambling валидиране

Статистическата оценка за статбилността на модела се получав чрез ниска средна стойноста на коефициента на определяне:

$$R^2_{YS} = \frac{1}{500}\sum_{k=1}^{500} R^2{}_{(k)}$$

За стабилен модел се изисква $R^2_{YS} < 0.10$

Ниските стойности на R^2_{YS} показват, че в оригиналния модел няма случайни корелации между дескрипторите и целевото свойство.

Литература:

1. Chemoinformatics: a textbook, Ed. J. Gasteiger & T. Engel, WILEY-VCH Verlag GmbH & Co., 2003.

2. Handbook of Chemoinformatics: From Data To Knowledge Vol. 3, Ed. J. Gasteiger, WILEY-VCH Verlag GmbH & Co., 2003.

3. D. Massart, B. Vandenginste, S. Deming, Y. Michotte, L. Kaufman. Chemometrics: a textbook. Elsevier, 1988.