

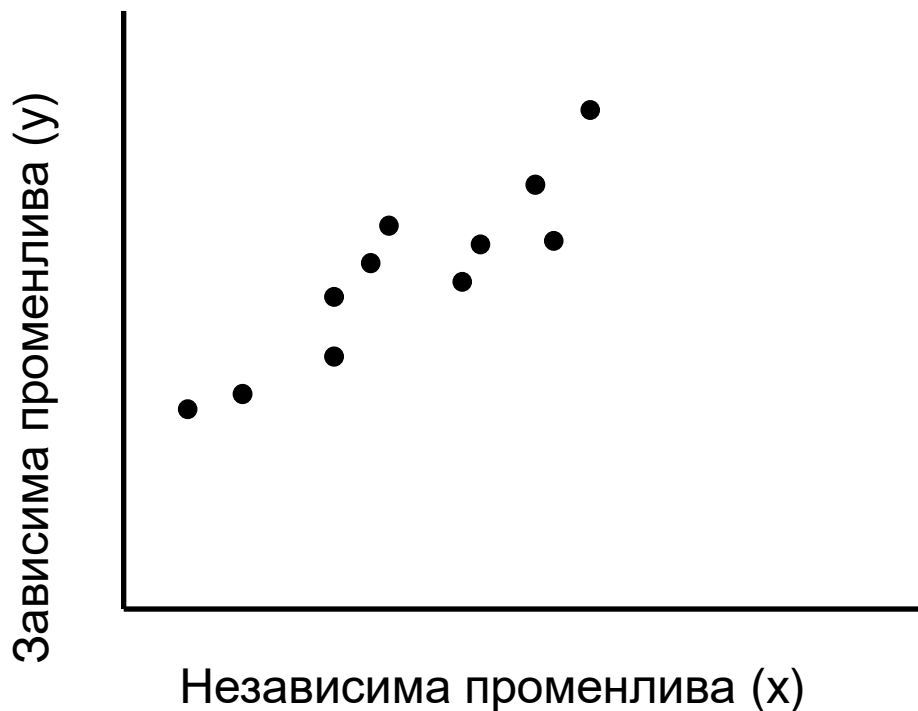
# Изкуствен интелект

*Тема 16: Методи за машинно обучение. Регресия. Линейна регресия. Анализ на грешката.*

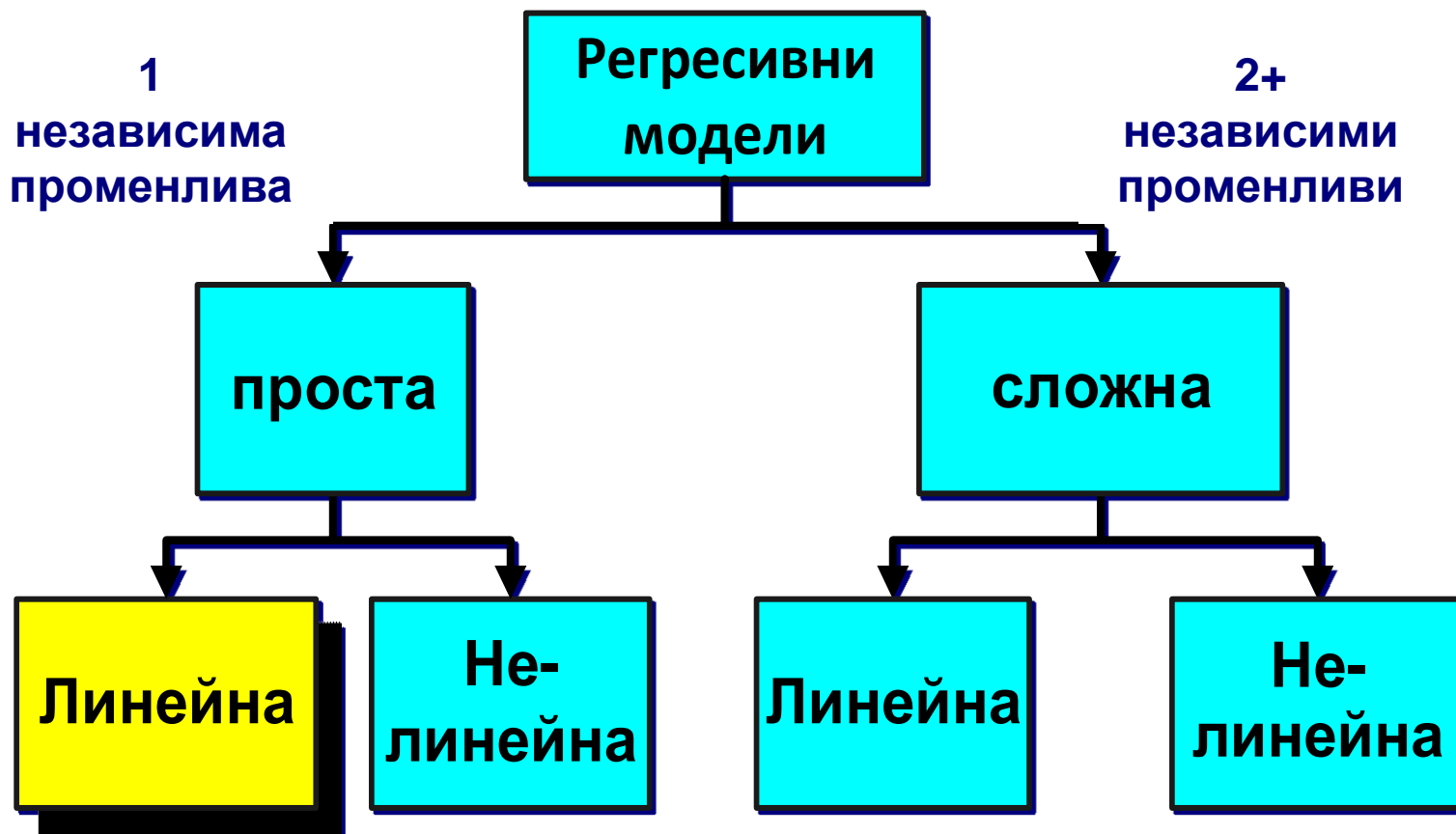


# Регресия, регресионен анализ

- ❑ Регресията цели да обясни (моделира) изменението на зависимата променлива използвайки знание за изменението на една или повече независими променливи.
- ❑ Регресията е следствие от каузалността.
- ❑ Ако независимите променливи са свързани с изменението на зависимата променлива, моделът може да се използва за предсказване...



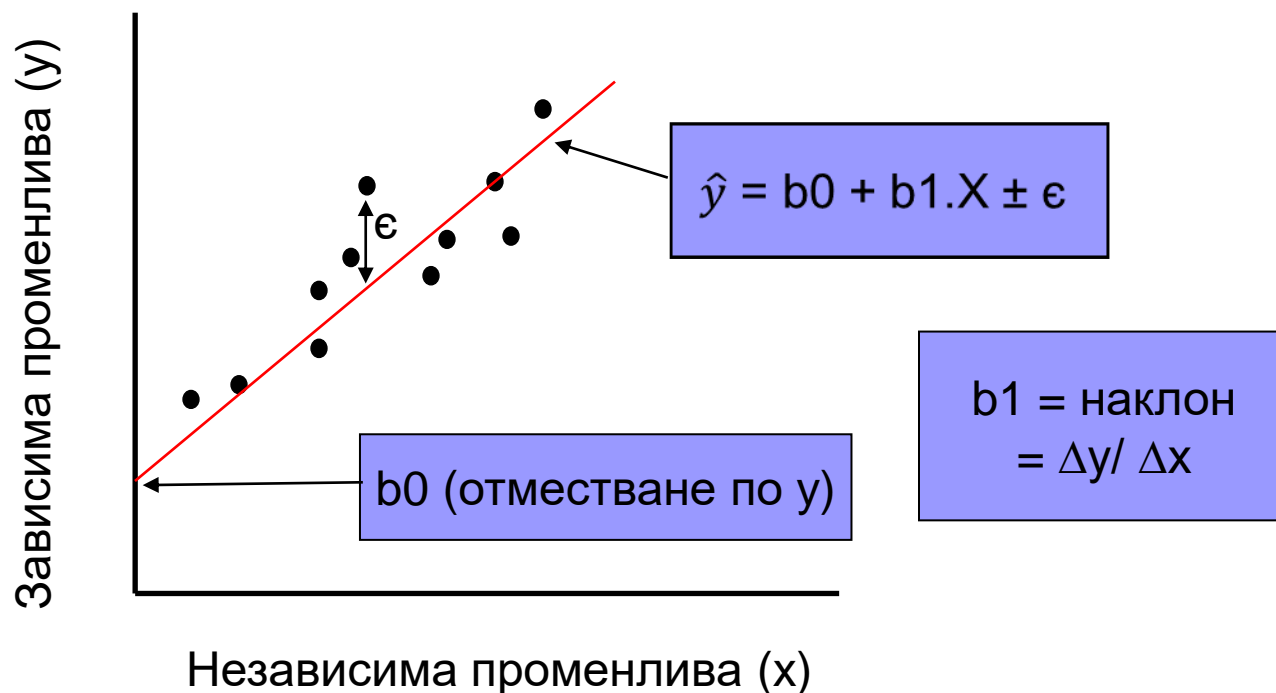
# Регресия, регресионен анализ



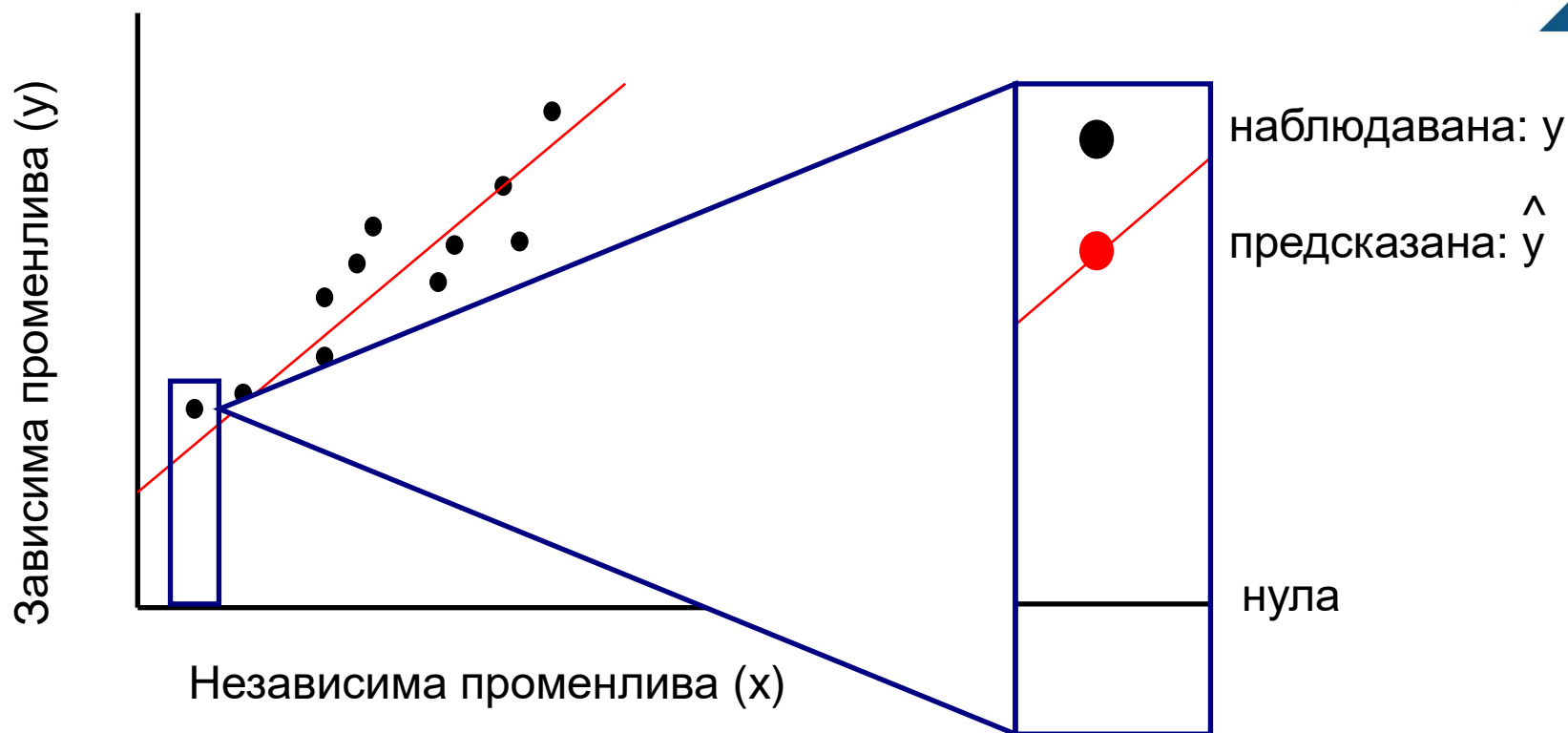
# Проста линейна регресия

Резултатът от регресията е функцията, която предсказва зависимата променлива въз основа на стойностите на независимата променлива.

При простата линейна регресия апроксимираме с помощта на права линия.



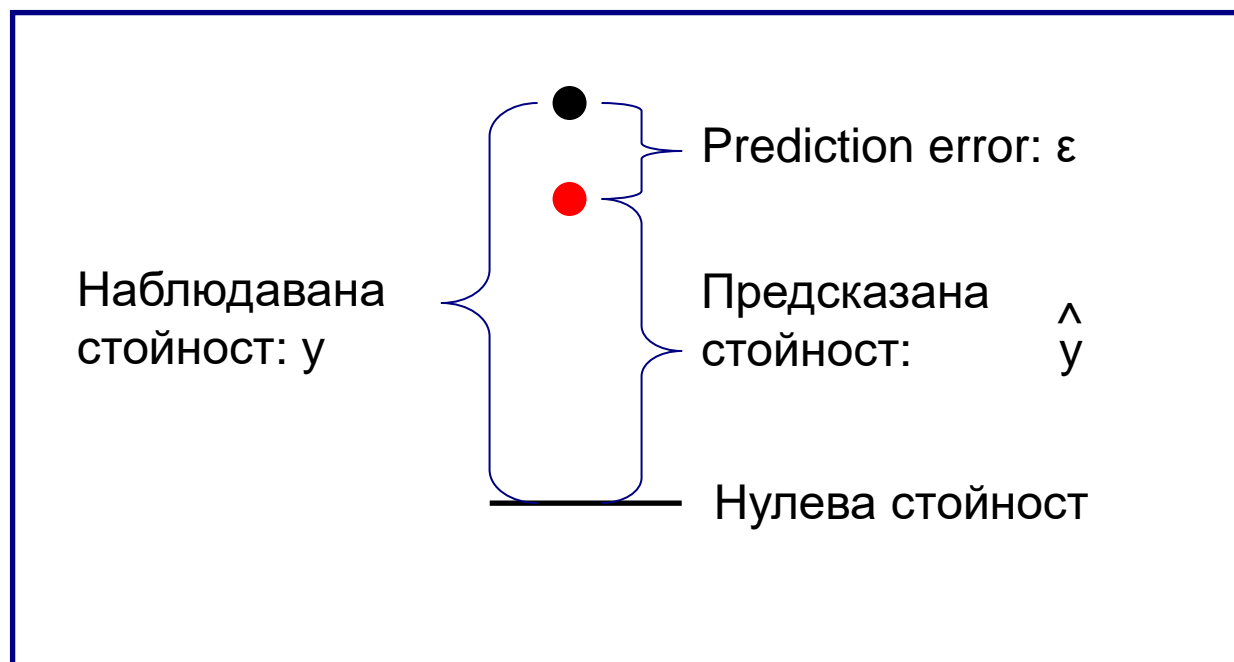
# Проста линейна регресия



Функцията трябва да предскаже ст-та на зависимата променлива  $y$  по зададена ст-ст на  $x$

Наблюдаваната ст-ст е означена с  $y$ , а предсказаната с  $\hat{y}$ .

# Проста линейна регресия

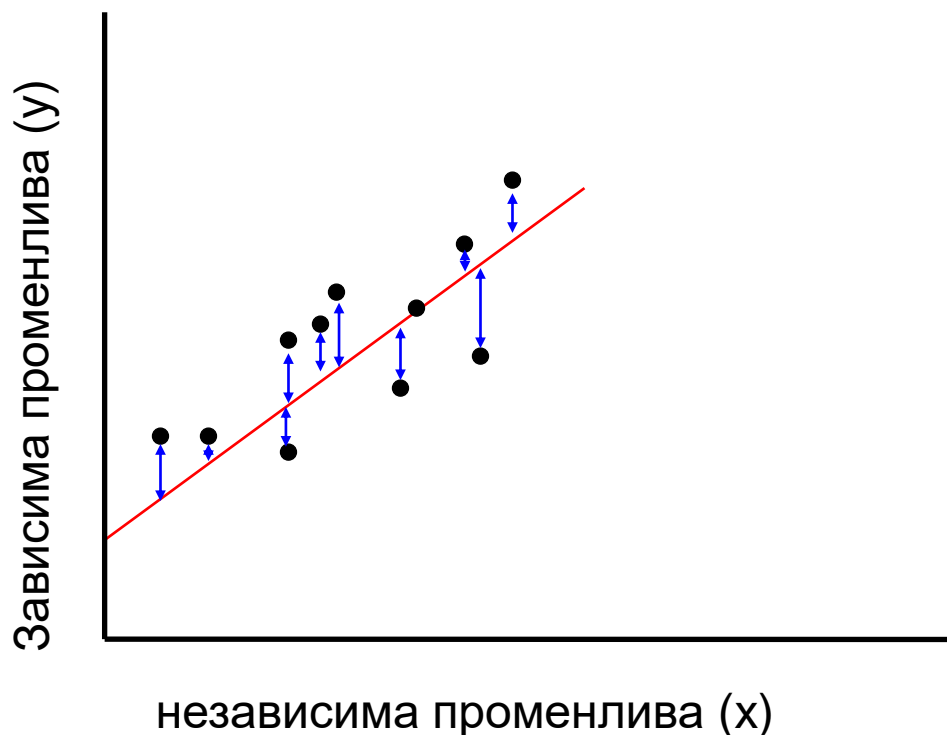


За всяко наблюдение, отклонението може да бъде означено като:

$$y = \hat{y} + \epsilon$$

истинска = обяснена + грешка

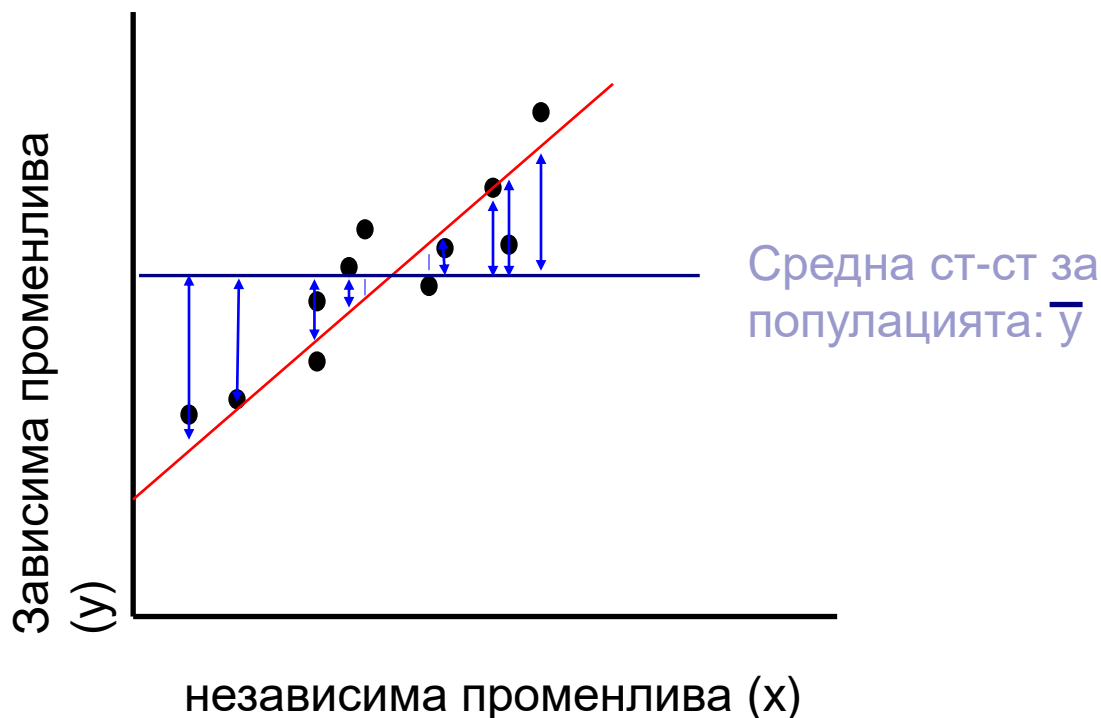
# Регресия



По метода на най-малките квадрати може да се избере линия която да доведе до най-малката стойност на сумата от квадрата на грешките на предсказване.

Тази стойност се нарича сума от квадратите на грешката (Sum of Squares of Error, SSE).

# Регресия



Регресия със сума от квадратите -- сумата от квадратите на разликите м/у предсказаната стойност на всяко наблюдение и средната стойност на популацията.



# Регресия

Обща сума на квадратите (Total Sum of Squares, SST) е

$$SST = SSE + SSR$$

Total sum  
of Squares

Sum of  
Squares Error

Sum of Squares  
Regression

$$SST = \sum (y - \bar{y})^2$$

мярка за общата  
изменчивост на  $y$

$$SSE = \sum (y - \hat{y})^2$$

мярка за необяснената  
от  $\hat{y}$  промяна

$$SSR = \sum (\hat{y} - \bar{y})^2$$

мярка за обяснимата  
промяна

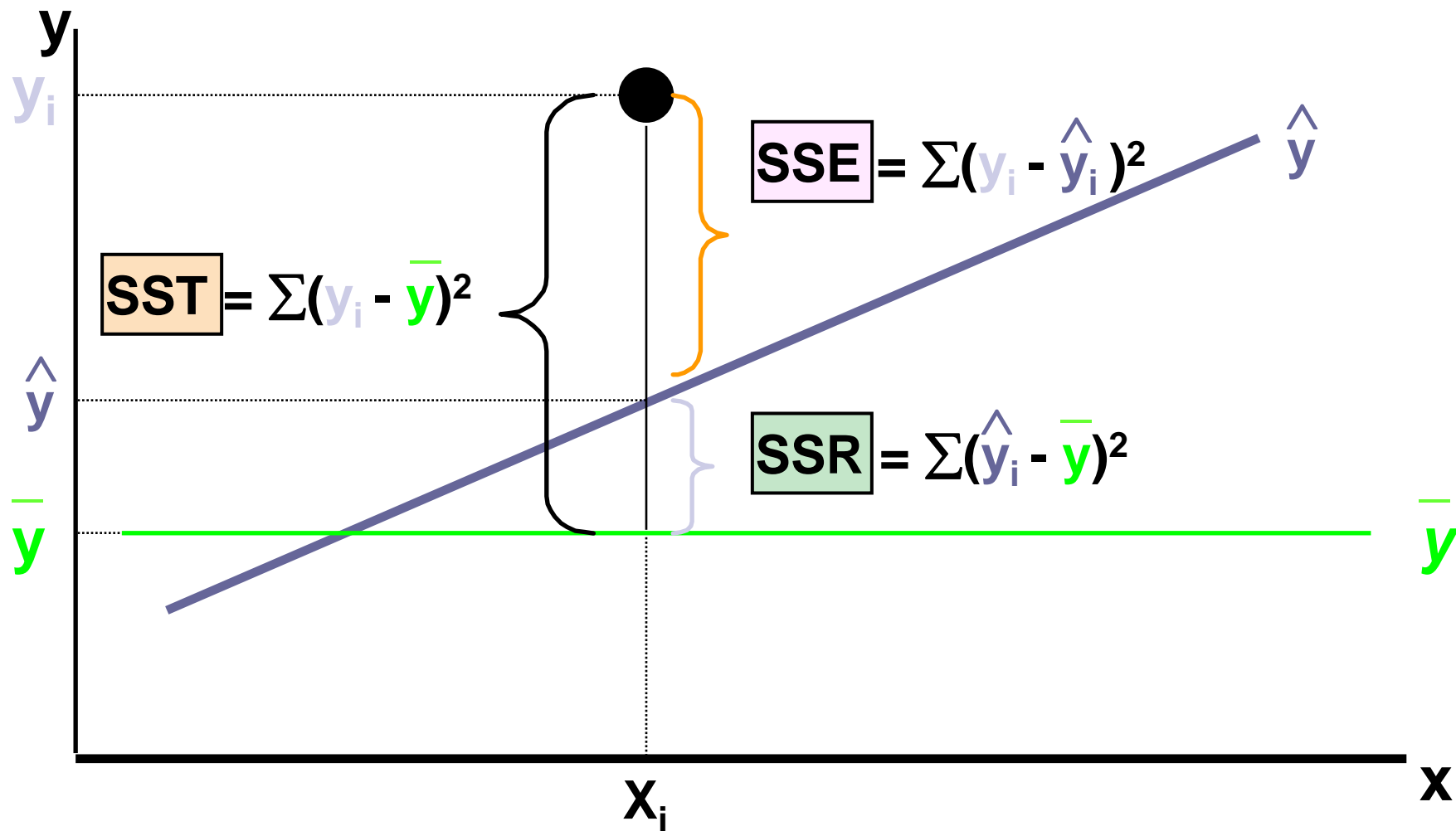
where:

$\bar{y}$  = средна ст-ст за зависимата променлива

$y$  = наблюдавана стойност

$\hat{y}$  = очаквана стойност на  $y$  при зададено  $x$

# Регрессия



# Регресия

Пропорцията на общата изменчивост (SST) отнесена към мярка за обяснимата промяна чрез регресията (SSR) се нарича коефициент на определеност (Coefficient of Determination), често се означава с  $R^2$ .

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} \quad 0 \leq R^2 \leq 1$$

Стойността на  $R^2$  е м/у 0 и 1, и колкото е по-висока, толкова по-точен е модела на регресията.

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Оценката на достоверния интервал за отделна ст-ст на  $y$  при дадена определена ст-ст на  $x_p$

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

Този член добавен към ширината на интервала отговаря на добавената несигурност за всеки отделен случай

# Стандартна грешка при регресията

- Стандартна грешка при регресията е мярката за нейната вариативност и може да се използва в смисъла на стандартното отклонение, за предсказване на интервалите на достоверност.
- $y \pm 2$  стандартни грешки осигурява прецизност от приблизително 95%, а 3 стандартни грешки осигуряват доверителен интервал от 99%.
- Стандартната грешка се определя като корен квадратен от средната грешка на предсказване.

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n - k - 1}}$$

$n$  е броя наблюдения в извадката

$k$  е общия брой на променливите в модела

# Стандартна грешка при регресията

Стандартната грешка на коефициента на наклон на регресията ( $b_1$ ) е

$$S_{b_1} = \frac{S_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{S_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

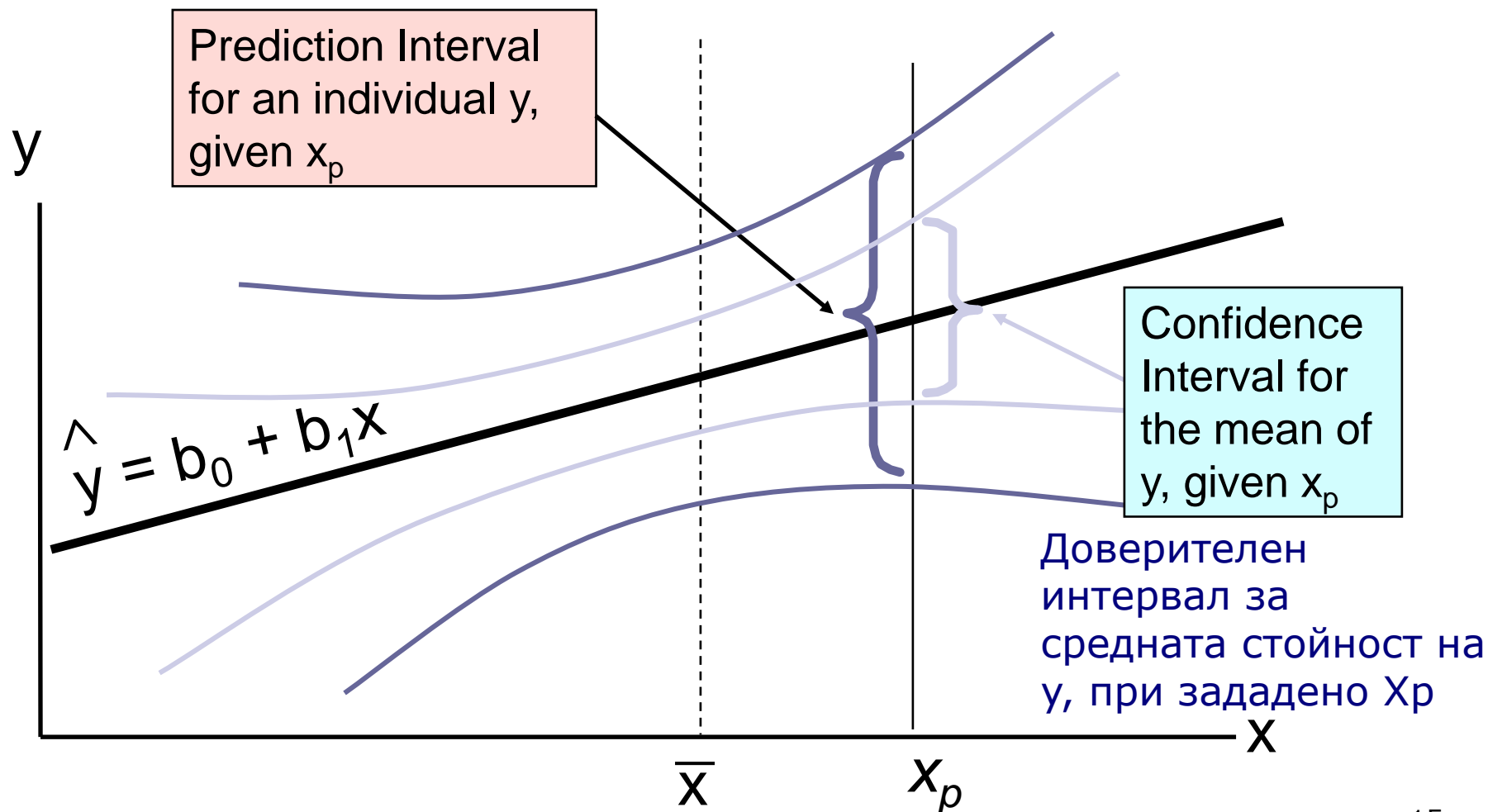
където:

$S_{b_1}$  = оценка на стандартната грешка по метода на най-малките квадрати

$S_\varepsilon = \sqrt{\frac{SSE}{n-2}}$  = оценка на стандартната грешка за извадката

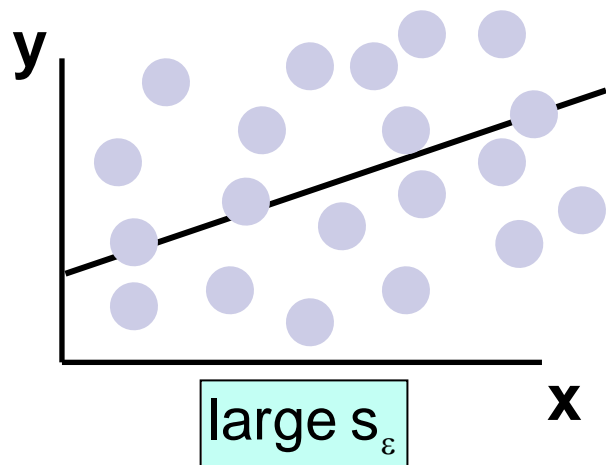
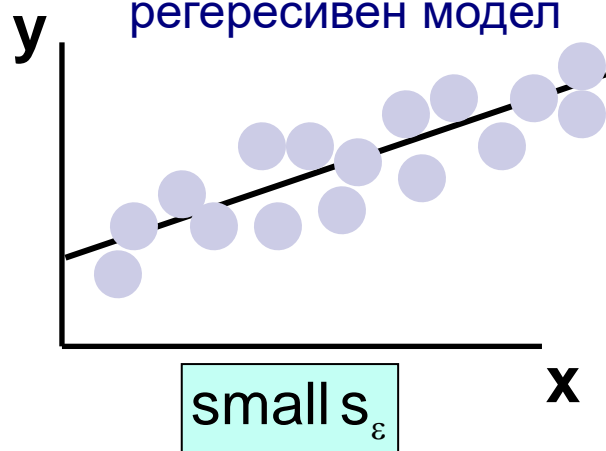
# Интервална оценка за различни ст-ти на $x$

Интервал на прогнозиране за конкретна стойност на  $y$ , при зададена ст-ст  $x_p$

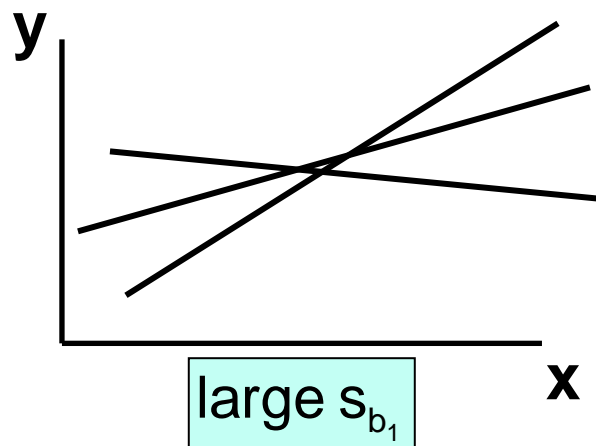


# Стандартна грешка при регресията

Отклонение от линейния  
регресивен модел



Изменчивост в наклона на  
модела спрямо ст-стите





# Анализ на грешката

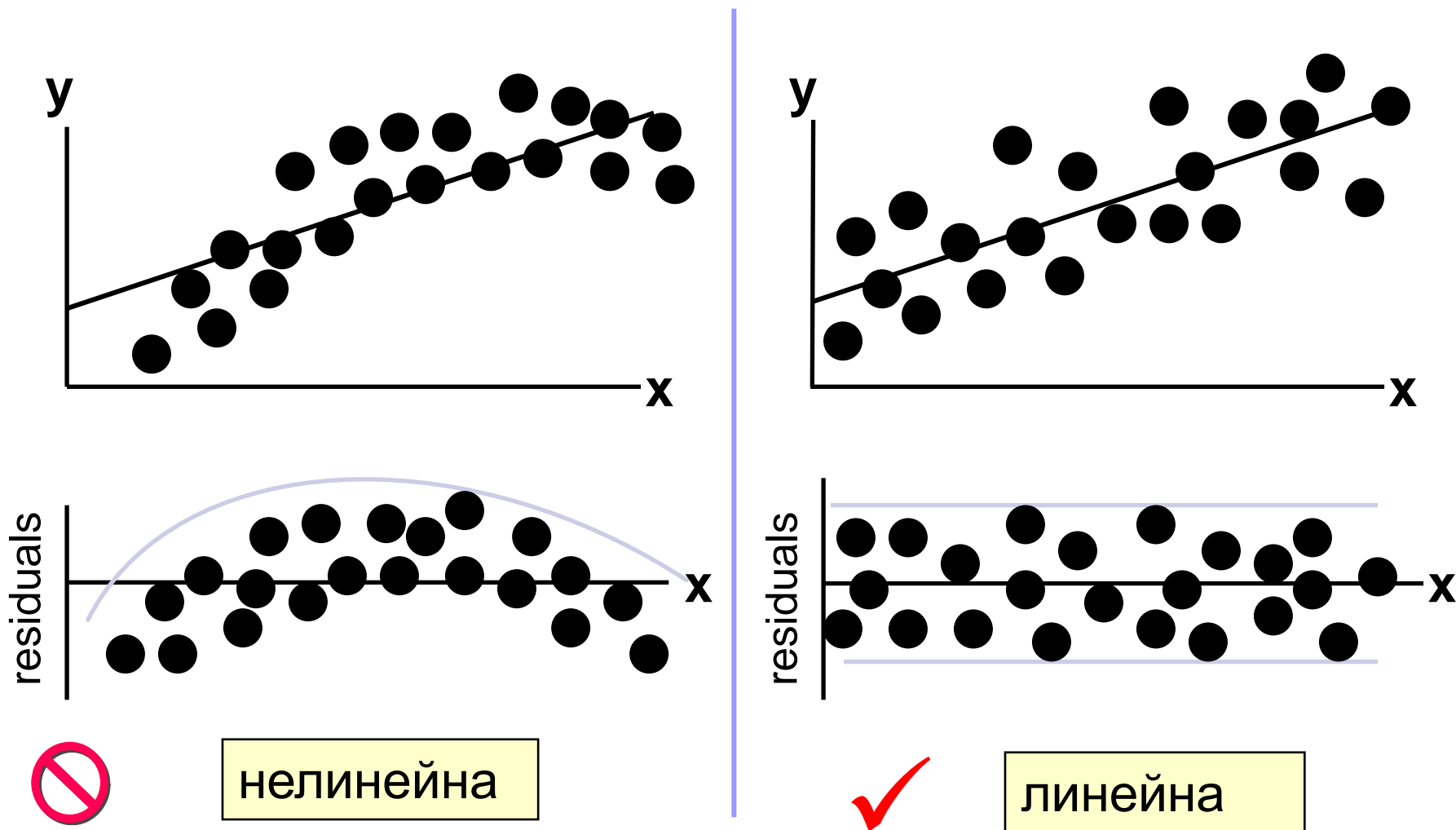
## ■ Цел

- ☐ За да се провери допускането за линейност
- ☐ Да провери дали стандартното отклонение за всяка стойност на  $x$  дали е еднаква
- ☐ Да се оцени допускането за нормалност на разпределението

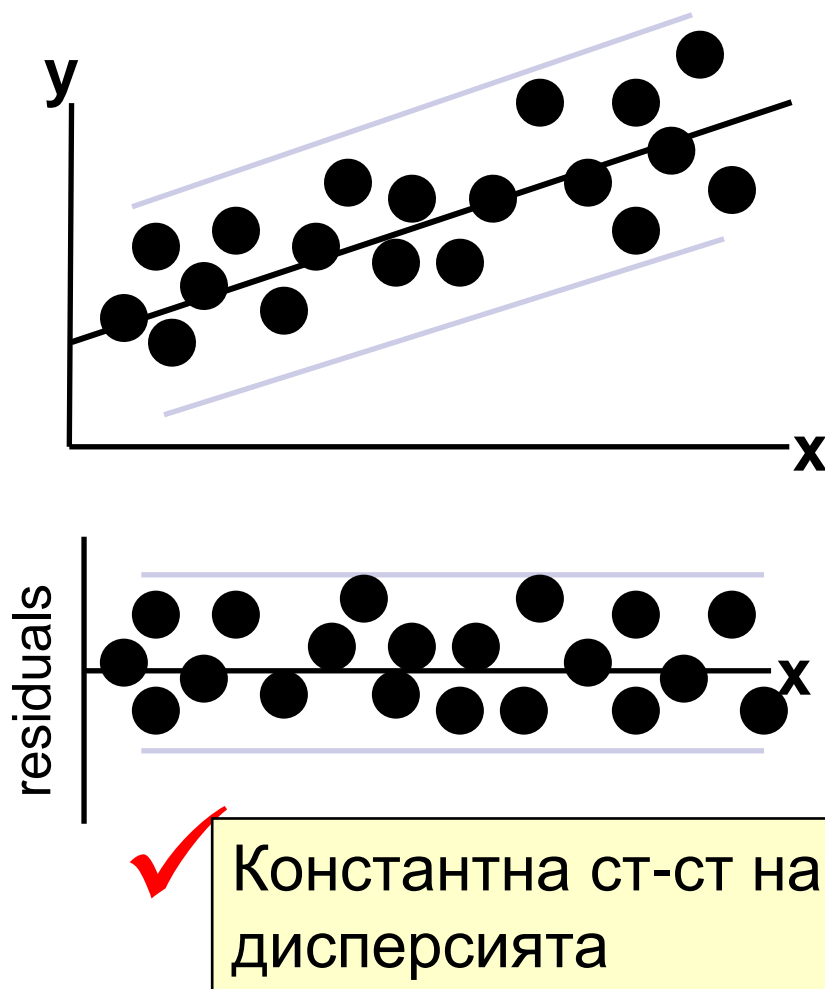
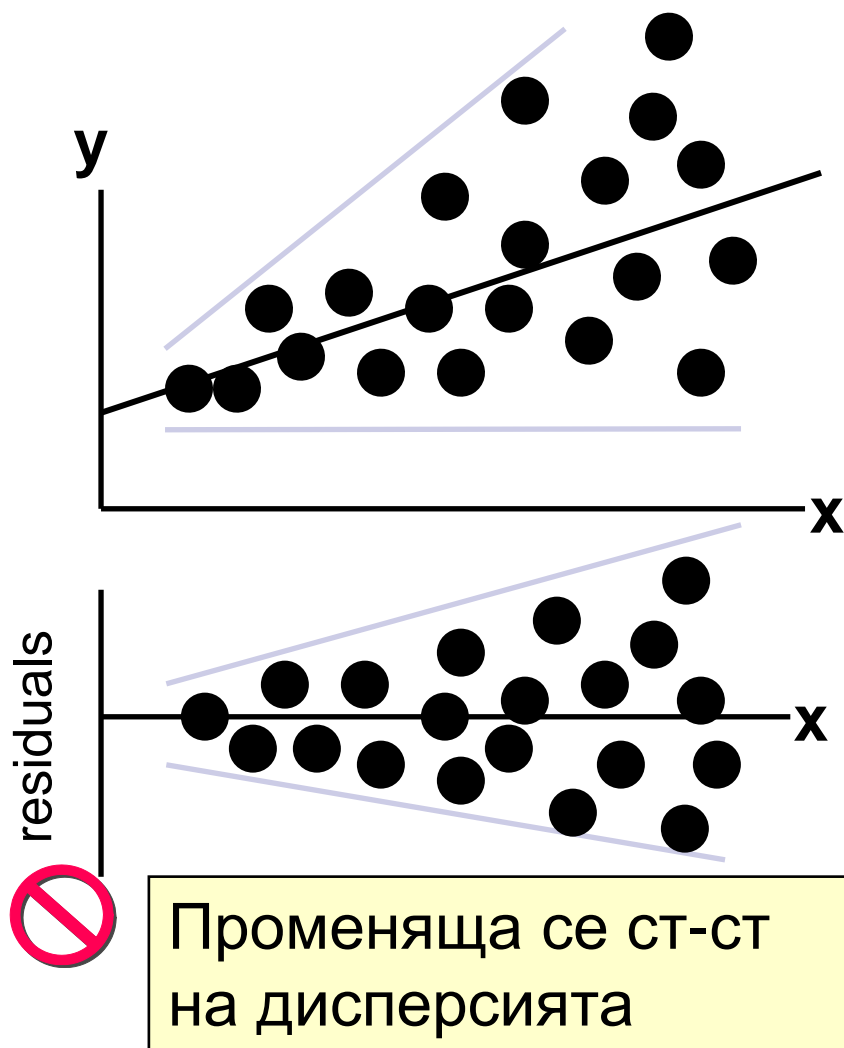
## ■ Графичен анализ на грешката

- ☐ Можем да изобразим грешката спрямо  $x$
- ☐ Можем да построим хистограма на грешката за да проверим за нормалност

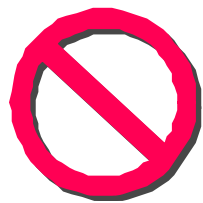
# Анализ на грешката за линейност



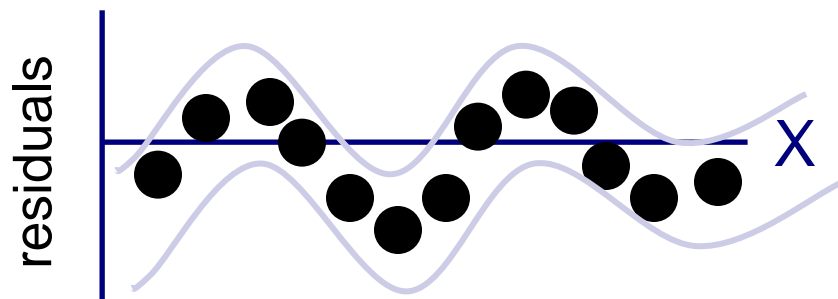
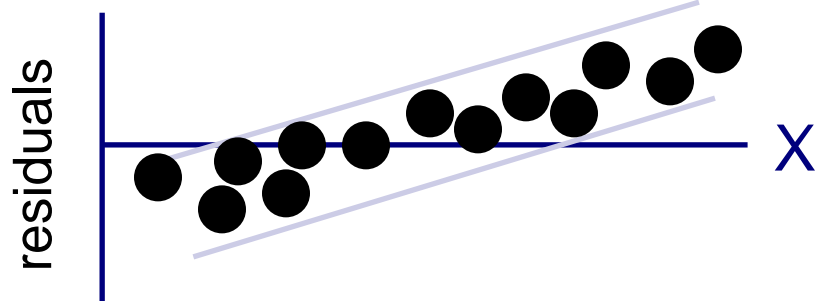
# Анализ на грешката за константно отклонение



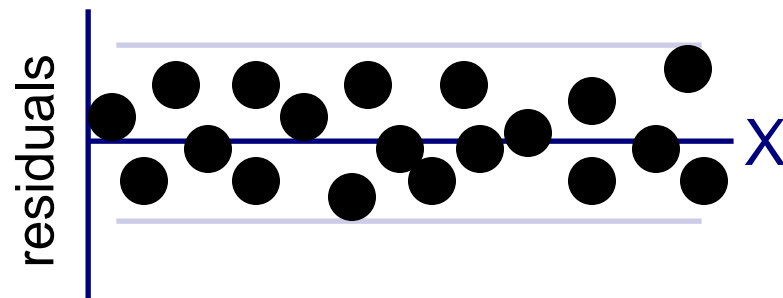
# Анализ на грешката за независимост



Грешката е ф-я на  $X$

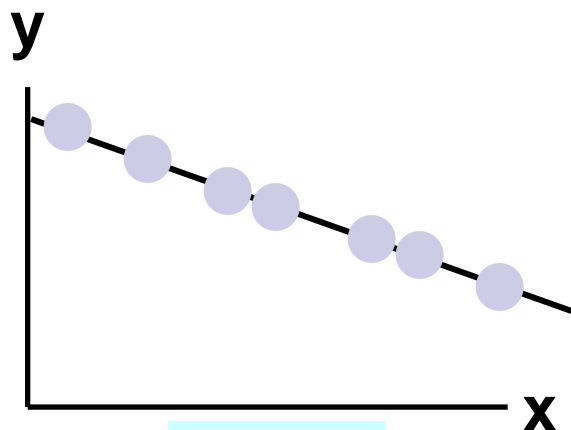


Независима от  $X$

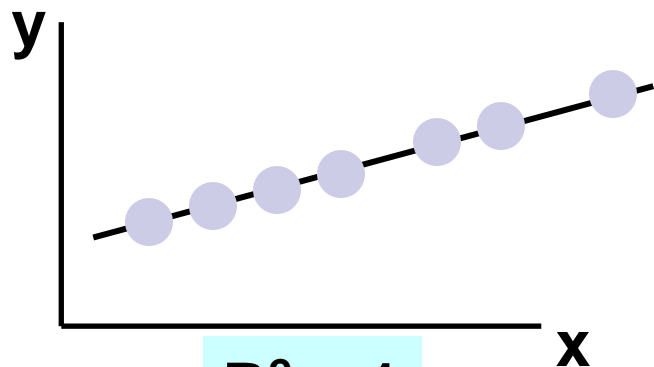
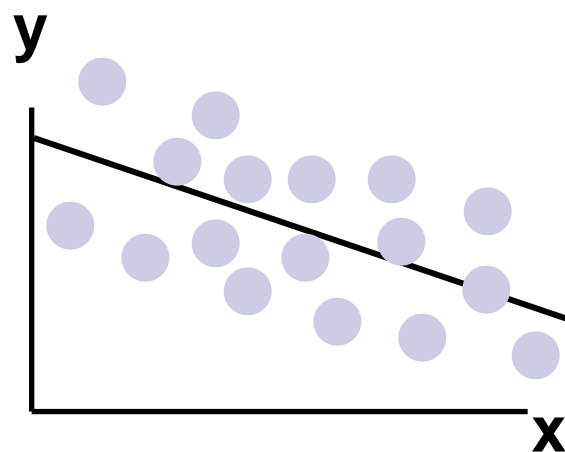


# Примери

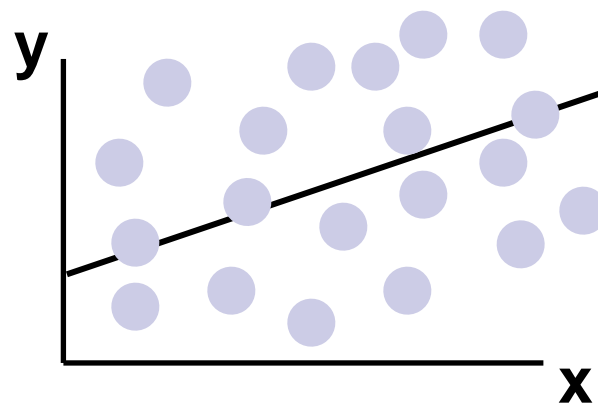
$$0 < R^2 < 1$$



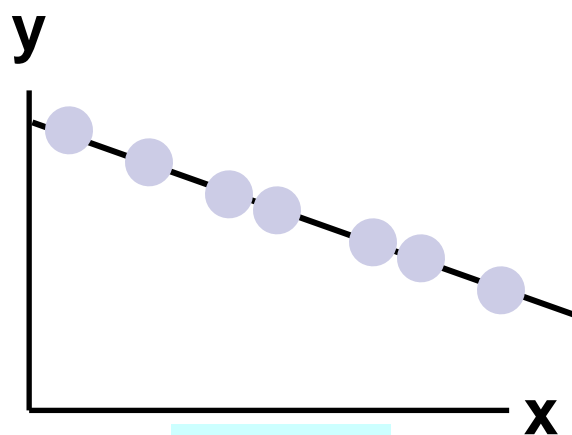
$$R^2 = 1$$



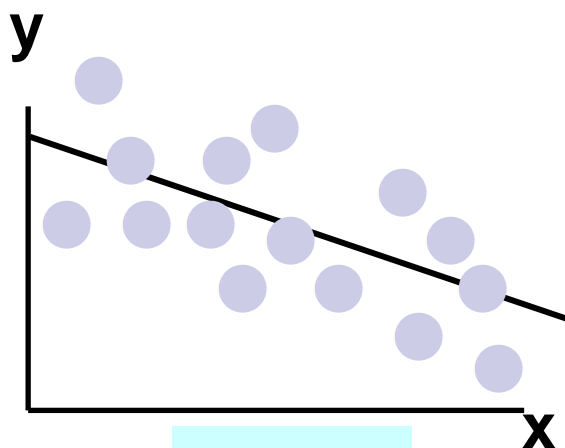
$$R^2 = 1$$



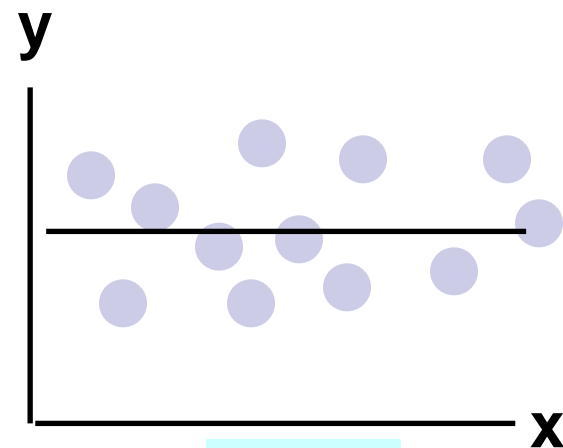
# Примери



$$r = -1$$



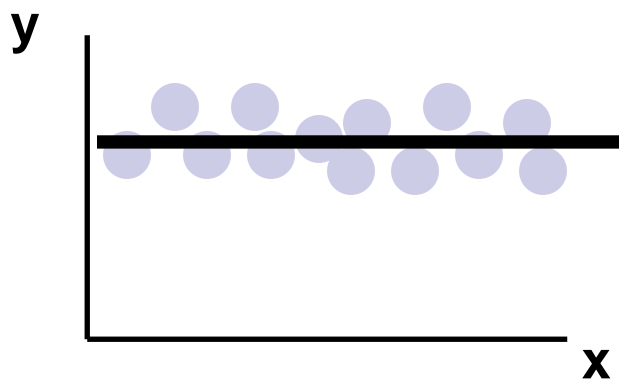
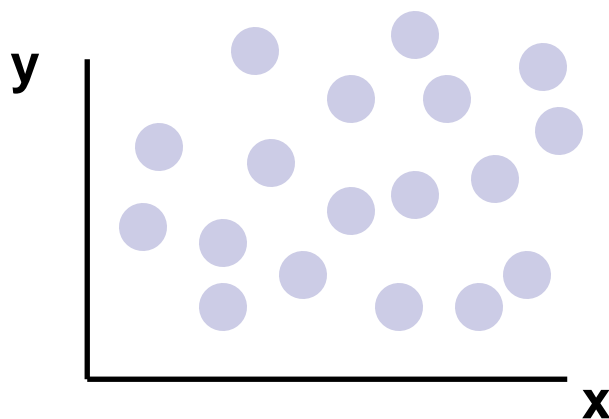
$$r = -.6$$



$$r = 0$$

# Примери

$$R^2 = 0$$



# Проста линейна регресия

Резултатът от проста линейна регресия е изчислена стойност на коефициента  $\beta$  и константата  $A$ :

$$y = A + \beta * x + \varepsilon$$

където  $\varepsilon$  е остатъчната грешка.

$\beta$  е изменението на зависимата променлива при единично изменение на независимата променлива:

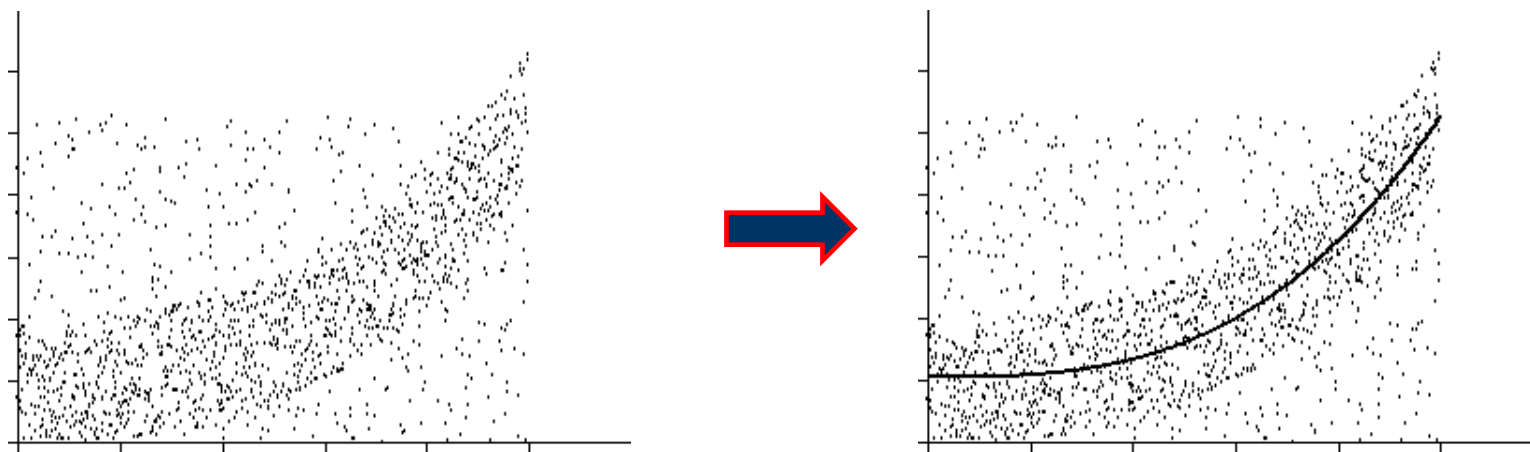
$$\beta = \frac{\Delta y}{\Delta x}$$

$$\hat{\beta} = \frac{Cov(x, y)}{Var(x)}$$

$$\text{Calculate : } \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$



# Нелинейна регресия



Нелинейна функция също може да бъде апроксимираща функция на регресията – например квадратична функция, логаритмична функция, експоненциална функция, или всяка друга непрекъсната функция.

# Множествена линейна регресия

Множествена линейна регресия - когато има повече от една независима променлива, която може да се използва за да се обясни изменението на зависимата променлива, но с ограничението независимите променливи да не са линейно свързани.

Множествена линейна регресия се дефинира като

$$y = A + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

където  $k$  е броя на променливите или параметрите на модела.

# Мультиколинеарност

Мультиколинеарност е условието при което най-малко 2 независими променливи са линейно корелирани.

*Example table of  
Correlations*

	Y	X1	X2
Y	1.000		
X1	0.802	1.000	
X2	0.848	0.578	1.000

Корелационната таблица подсказва кои независими променливи са свързани.

В общия случай, независимата променлива която има ст-ст на корелацията по-голяма от 0.3 със зависимата променлива и по-малко от 0.7 с всяка друга независима променлива която може да се използва като предсказател.