REGRESIJSKA ANALIZA

<u>REGRESIJSKA ANALIZA</u>

• često imamo dvije ili više varijabli koje su inherentno povezane, odnosno postoji neka **zavisnost (korelacija)** među njima koju želimo istražiti

• **regresijske tehnike** omogućuju nam da **kvantitativno** izrazimo takvu **zavisnost (korelaciju)** te dobiveni model koristimo za (1) <u>predviđanje nekih podataka za koje nemamo mjerenja</u> ili ga koristimo da dođemo do nekih (2) <u>konstanti koje nam tu zavisnost opisuju</u> i sl. (*npr*. zavisnost temperature i brzine kemijske reakcije)

• u najjednostavnijem slučaju imamo linearnu zavisnost jedne varijable *(y)* o jednoj nezavisnoj varijabli *(x)* – ***linearna regresija***

• moguća je i zavisnost jedne varijable (*y*) o više nezavisnih varijabli – **multivarijatna regresija**

• **nelinearna regresija**

REGRESIJSKA ANALIZA

• lat. *regressio - povrat, vraćanje, odstup, uzmak, nazadovanje*

• Sir Francis Galton, znanstvenik iz 19. stoljeća - koncept korelacije i regresije

• metoda koja proučava ovisnost između varijabli i predstavlja najčešće upotrebljavanu statističku metodu **(regresija ili interpolacija)**

• linearna ovisnost je najjednostavnija netrivijalna ovisnost koja se može zamisliti

• prava ovisnost između varijabli je često približno linearna u nekom podskupu vrijednosti neovisne varijable

• često se **nelinarna ovisnost** između varijabli može matematičkim operacijama prevesti u linearnu ovisnost, ali postoji i **nelinearna regresija**

REGRESIJSKA ANALIZA

- **neovisna varijabla** je varijabla čiju vrijednost određuje osoba koja provodi pokus
- **ovisna varijabla** je varijabla čije vrijednosti ovise o vrijednosti neovisne varijable
- u praksi se za **neovisnu varijablu** uzima ona fizikalna veličina koja se može **najtočnije mjeriti**

- pretpostavka je da postoji funkcija $f(x)$ takva da se za svaku vrijednost neovisne varijable x_i , **ovisna varijabla** može napisati kao: $\boldsymbol{y_i = f(x_i) + e_i}$ gdje je pogreška $\boldsymbol{e_i}$ slučajna varijabla s **normalnom raspodjelom** i **očekivanom vrijednošću nula**

- npr. linearna regresija - pretpostavka je da postoje koeficijenti a i b takvi da se za svaku vrijednost neovisne varijable x_i , ovisna varijabla može napisati kao:

$\boldsymbol{y_i = bx_i + a + e_i}$

- koeficijenti a i b najčešće se određuju **metodom najmanjih kvadrata** koja minimizira vrijednosti kvadrata udaljenosti između opaženih podataka i regresijske krivulje (pravca)

REGRESIJSKA ANALIZA

• matematički postupak za pronalaženje krivulje koja prolazi kroz zadani skup točaka uz **minimiziranje sume kvadrata odstupanja** zadanih točaka od te krivulje

• kvadrati odstupanja se koriste zbog toga što to omogućuje da se **rezidui** tretiraju kao **kontinuirana veličina**

• no korištenje kvadrata odstupanja daje veće težinske faktore za **točke koje jako odstupaju od linearnog modela** što u nekim slučajevima može **otežati interpretaciju rezultata** *("outlieri" – vidi kasnije)*

LINEARNA REGRESIJA

• imamo zavisnost jedne varijable *(y)* o jednoj nezavisnoj varijabli *(x)*

• pretpostavljamo da je *x* kontinuirana slučajna varijabla te da je ovisnost između *y* i *x*

linearna ovisnost: ***y = bx + a***

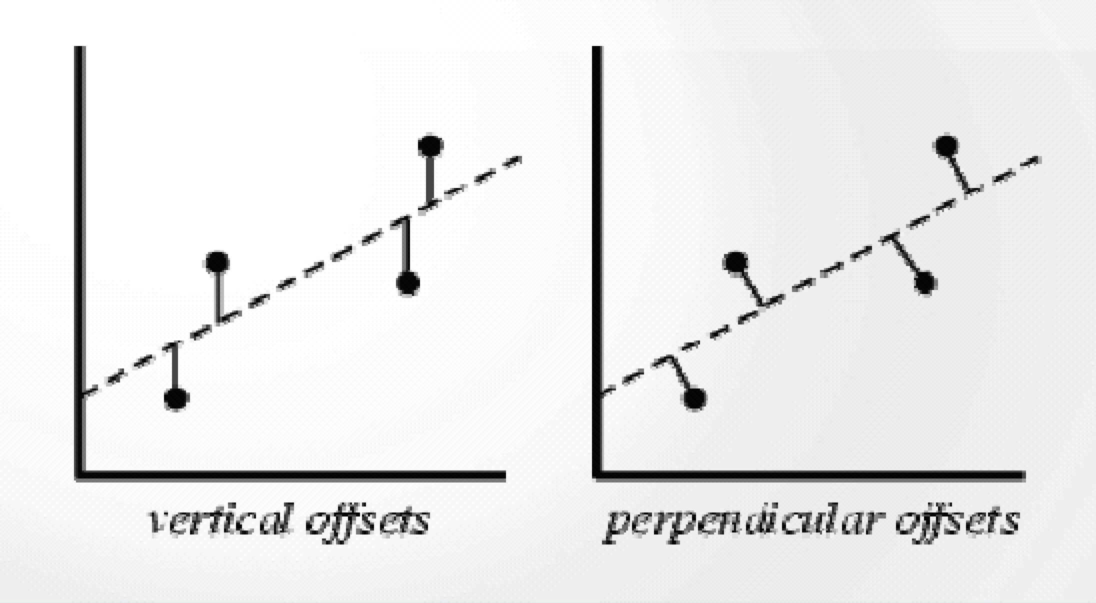
- IZVOD ZA IZRAZ METODE NAJMANJIH KVADRATA

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

$$\hat{a} = \frac{\overline{y}\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

LINEARNA REGRESIJA

• kod metode bazirane na **vertikalnim odstupanjima** sva slučajna **pogreška pripisana je *y* varijabli** (u takvim slučajevima se često u literaturi za *x* kaže “error-free ” varijabla)(npr. mjerimo brzinu reakcije u ovisnosti o koncentraciji i svu pogrešku pripišemo određivanju brzine)



• **nužan uvjet** za korištenje klasičnih regresijskih metoda je da broj parametara koje tražimo regresijom (*a* i *b* u slučaju pravca) bude **manji** (ili u najgorem slučaju **jednak**) od broja mjerenja (x_i, y_i) koje imamo na raspolaganju

LINEARNA REGRESIJA

r^2 – korelacijski koeficjent

- r – linearni korelacijski koeficijent
- govori o korelaciji i smjeru linearne povezanosti između dvije varijable
- +1 pozitivna korelacija; −1 negativna korelacija; 0 nema korelacije

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

- IZVOD I OBJAŠNJENJE

LINEARNA REGRESIJA

- koeficijent određivanja $0 \le r^2 \le 1$ predstavlja omjer **modelom opisanih varijacija** u odnosu na **ukupne varijacije podataka**

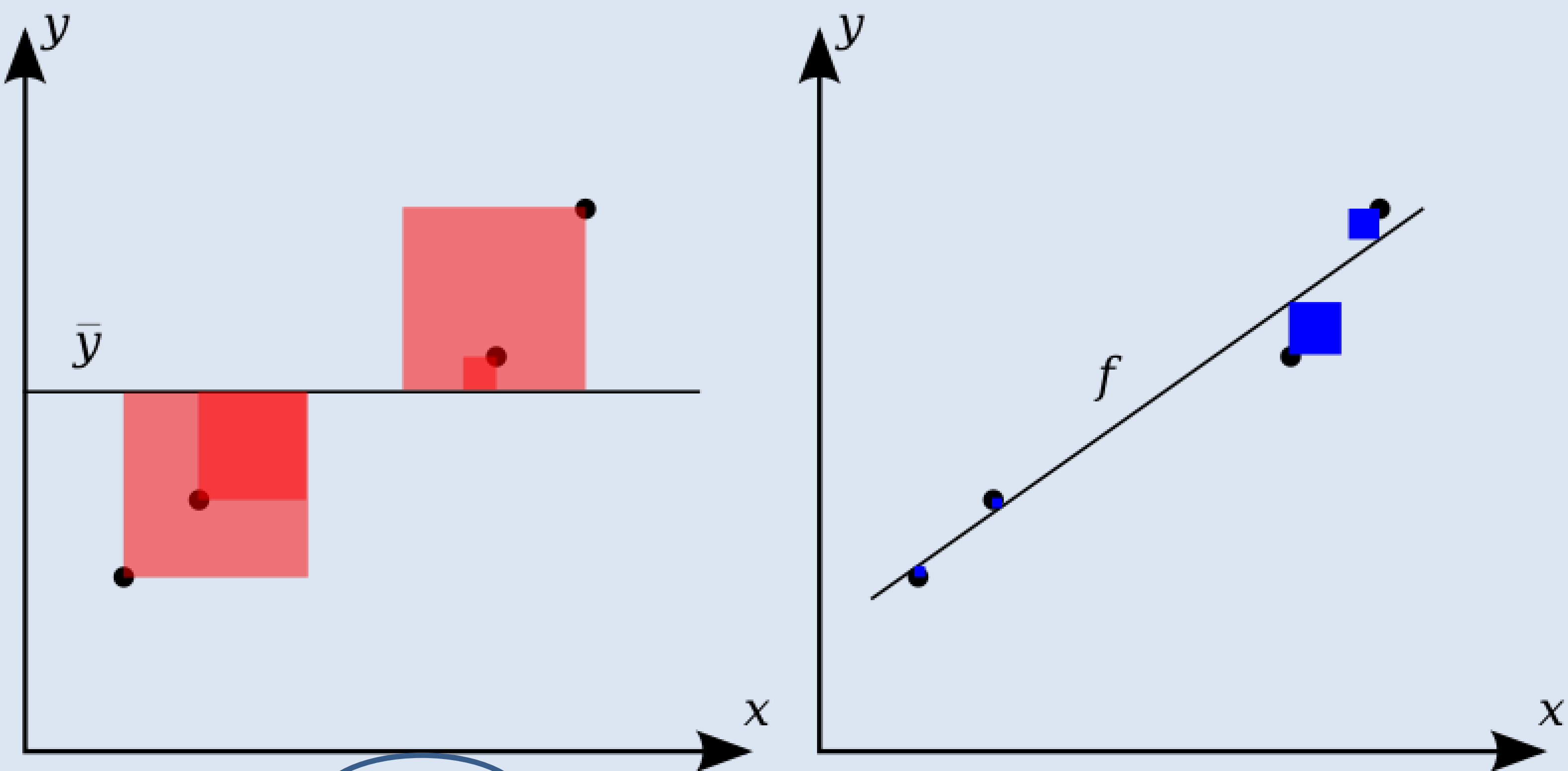
- daje **udio varijacije jedne varijable** koji se može predvidjeti iz druge varijable

- pokazuje koliko dobro linearna regresija opisuje zadane podatke

- ako je npr. $r = 0{,}9$ $r^2 = 0{,}81$ *to znači da se* 81 % ukupne varijacije u y može objasniti linearnom ovisnošću između x *i* y dok se preostalih 19 % ne može objasniti linearnom ovisnošću

- treba biti jako oprezan jer je r^2 lako "nafitati" na veću vrijednost (uvođenjem dodatnih parametara, dodavanjem varijabli, …), ali to ne znači da smo nužno popravili model i adekvatno opisali korelaciju između y i x

LINEARNA REGRESIJA

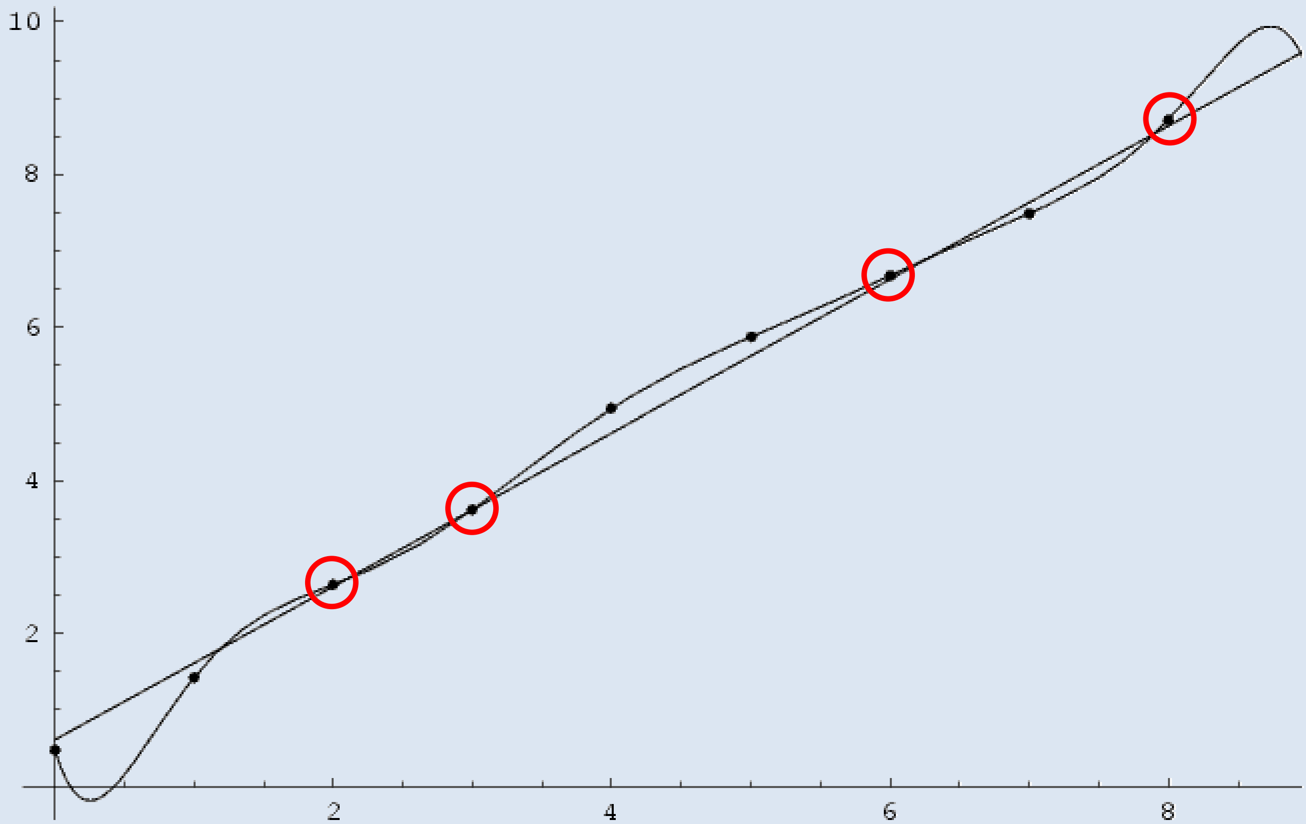


$$R^2 = 1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}}$$

SS**$_{tot}$* **– ukupna varijacija u vrijednostima *y

***SS**$_{err}$* **– udio varijacija u vrijednostima *y* koji nije objašnjen modelom** (ukoliko model valja pretpostavljamo da je nastao uslijed slučajne pogreške pri određivanju varijable *y*)

overfitting



LINEARNA REGRESIJA
Anscobeov kvartet (1973)

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

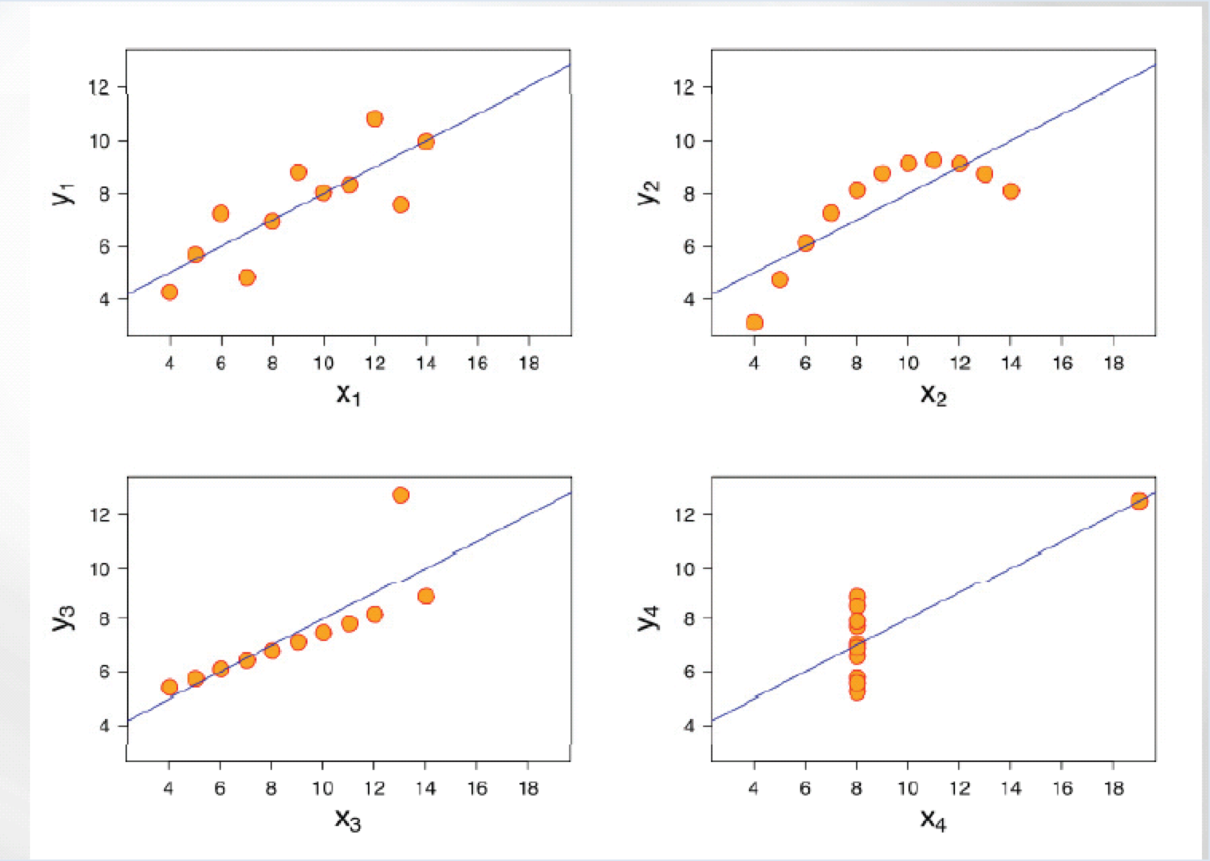
- sva četiri skupa podataka imaju istu vrijednost (najmanje na dvije decimale) za: ***r* = 0.816**

$\bar{x}$ = 9 $\bar{y}$ = 7,50 σ_x^2 = 11 σ_y^2 = 4,122 ili 4,127

$\boxed{y = 3.00 + 0.500x}$

LINEARNA REGRESIJA

Anscobeov kvartet (1973)



LINEARNA REGRESIJA

• **r^2** je statistik koji nam pokazuje koliko je slaganje između vrijednosti izračunatih modelom i izmjerenih vrijednosti, ali nam daje tek djelomične informacije o uspješnosti regresije u smislu objašnjavanja korelacije između zavisne i nezavisne varijable *(goodness of fit).*

• *goodness of fit* ("dobrota pristajanja") - koliko uspješno model opisuje **zavisnost (korelaciju)** između zavisne i nezavisne varijable

• vrlo često se koriste statistički testovi i hipoteze kako bi se izrazio *goodness of fit* (npr. χ^2-test, F-test, …)

MULTIVARIJATNA REGRESIJA

- imamo linearnu ovisnost jedne zavisne varijable *(y)* o više nezavisnih varijabli *(x_1, x_2, … x_n)*

$$y_i = a_1 \cdot x_{i1} + a_2 \cdot x_{i2} + \ldots + a_n \cdot x_{in}$$

- IZVOD: POJEDNOSTAVLJENI I/ILI PREKO SUME KVADRATA

- nužan uvjet: m > n (m je broj mjerenja koje imamo (broj točaka), n je broj parametara koje tražimo regresijom)

POLINOMNA REGRESIJA

• imamo polinomnu ovisnost zavisne varijable *(y)* o nezavisnoj varijabli *(x)*

$$y_i = a_1 \cdot x + a_2 \cdot x^2 + \ldots + a_m \cdot x^m$$

- IZVOD: POJEDNOSTAVLJENI I/ILI PREKO SUME KVADRATA

• polinomna regresija može se provesti na isti način kao i mulivarijatna regresija razmatrajući *x*, x^2, ... kao zasebne neovisne varijable

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\widehat{\vec{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\vec{y}.$$

• **nužan uvjet: m < n –** (ovdje smo zamijenili m i n; m je ovdje broj parametara koji regresijom tražimo, n je broj mjerenja koje imamo)

NELINEARNA REGRESIJA

• kod **linearne regresije**, prilikom korištenja metode najmanjih kvadrata javljali su se samo linearni oblici parametara a_i

• kod **nelinearne regresije**, prilikom korištenja metode najmanjih kvadrata javljaju se nelinearni oblici parametara a_i (poput a_i^2, e^{ai} i sl.)

•korištenjem metode najmanjih kvadrata dobije skup jednadžbi koje **nemaju jedinstveno rješenje** (naime, kod nelinearnih sistema derivacije sume najmanjih kvadrata su funkcije parametara i neovisne varijable te se ne mogu **jednoznačno riješiti**)

• za rješavanje takvih sustava potrebno je napraviti **početne procjene parametara** te ih zatim **iteracijskim postupkom utočniti**

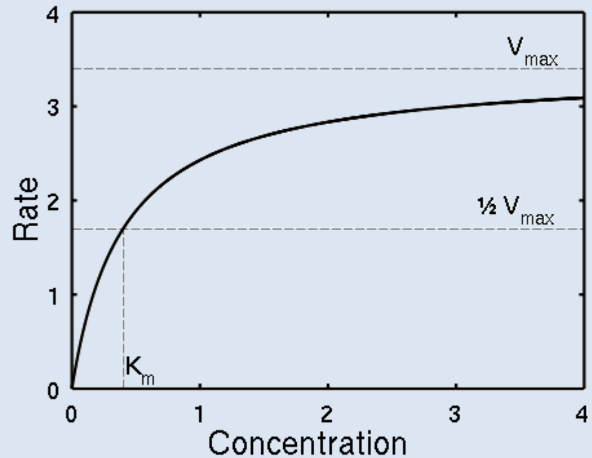
• koristimo neku od numeričkih metoda kako bi iteracijskim postupkom odredili parametre a_i pri tome je važno imati čim bolju početnu procjenu parametara te adekvatan iteracijski kriterij

• problemi: moguće je da imamo više rješenja, problem konvergiranja, …

NELINEARNA REGRESIJA

- u većini slučajeva se nelinerne zavisnosti nastoje linearizirati te provesti linearna regresija (*npr. Michaelis–Menten*)

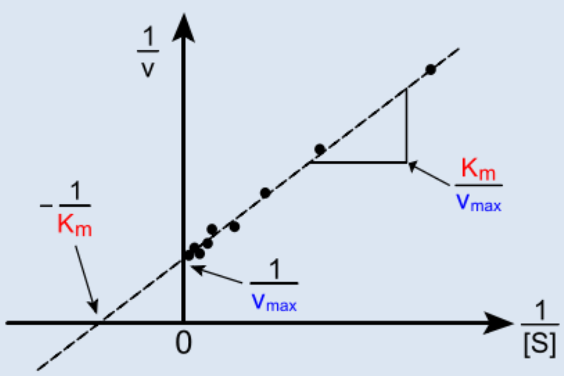
$$v = \frac{V_{\max}\,[\mathrm{S}]}{K_m + [\mathrm{S}]}$$

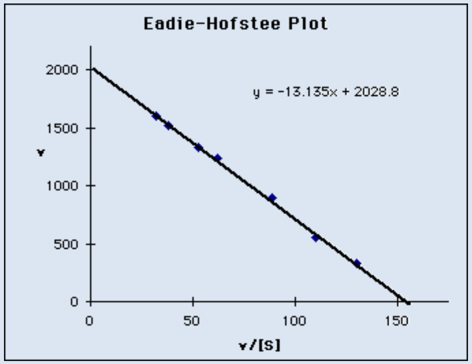


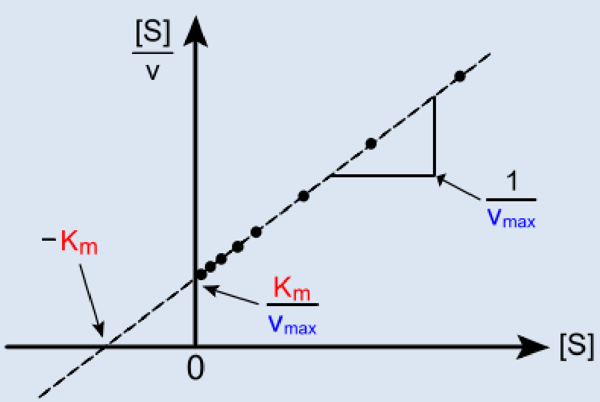
$$\frac{1}{v} = \frac{K_M}{V_{\max}[\mathrm{S}]} + \frac{1}{V_{\max}}$$

$$v = -K_m \frac{v}{[S]} + V_{\max}$$

$$\frac{[S]}{v} = \frac{1}{V_{\max}}[S] + \frac{K_m}{V_{\max}}$$







PLS *(Partial Least Square)* - metoda parcijalnih projekcija najmanjih kvadrata

PCA *(Principal Component Analysis)* – analiza glavnih komponenenata