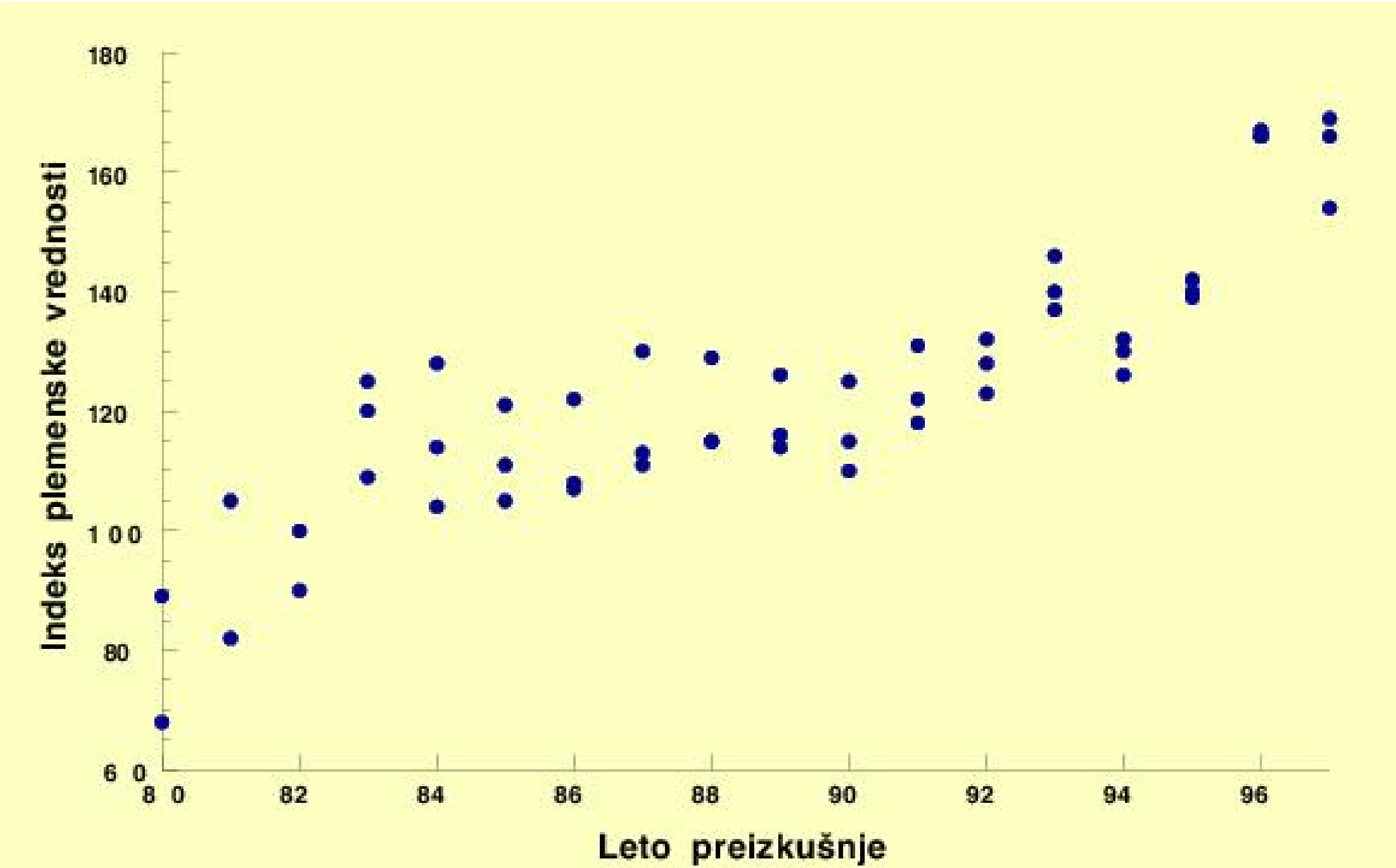
Statistični modeli
- regresija -

Milena Kovač

16. november 2012

Linearna regresija



Sestavimo model!

Neodvisna spremenljivka: leto preizkušnje (x, kvantitativna in diskretna spremenljivka, regresija)

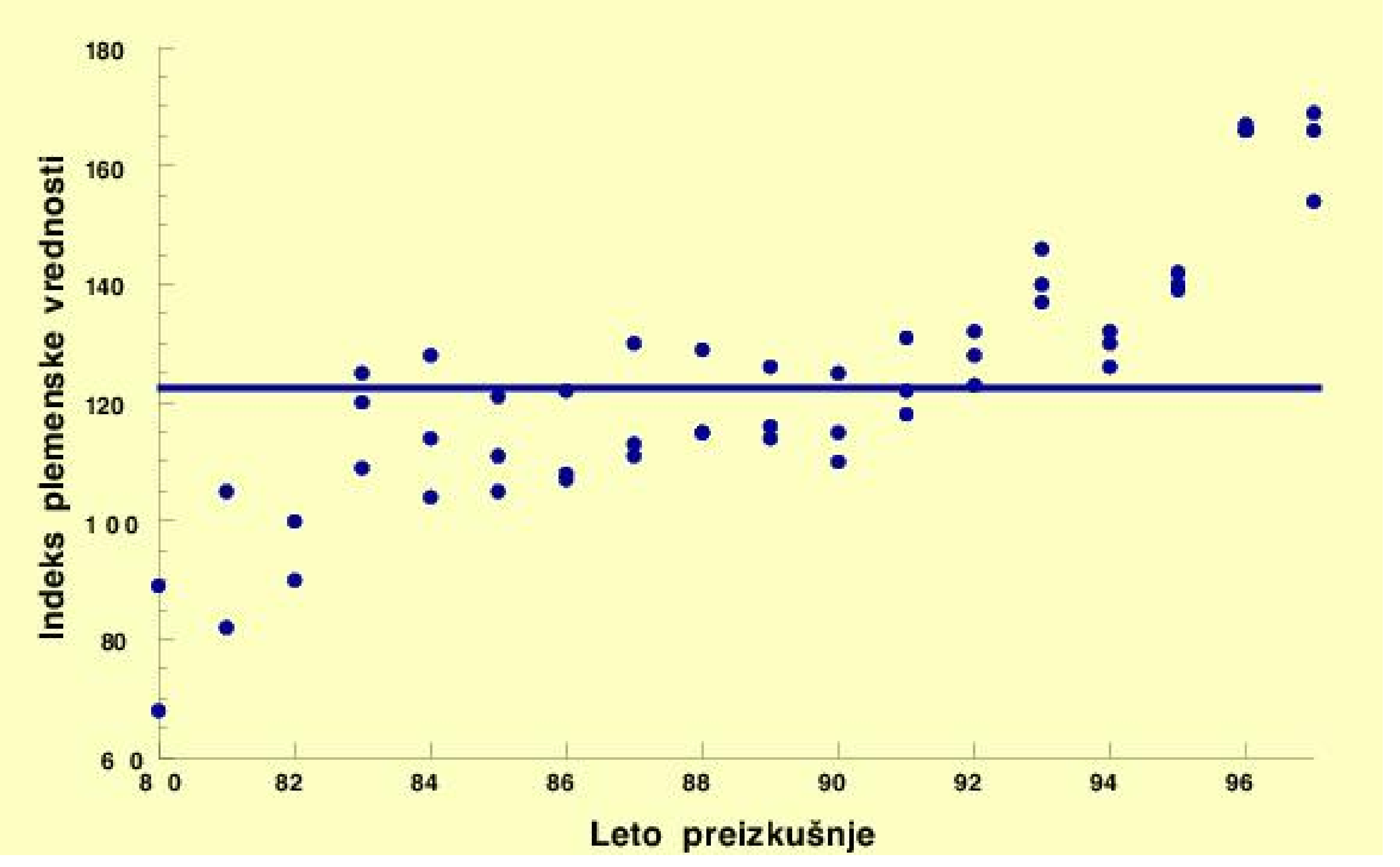
Odvisna spremenljivka: indeks plemenske vrednosti (y, kvantitativna in zvezna spremenljivka, N-porazdelitev)

Model: $y_i = \mu + b(x_i - 80) + e_i$

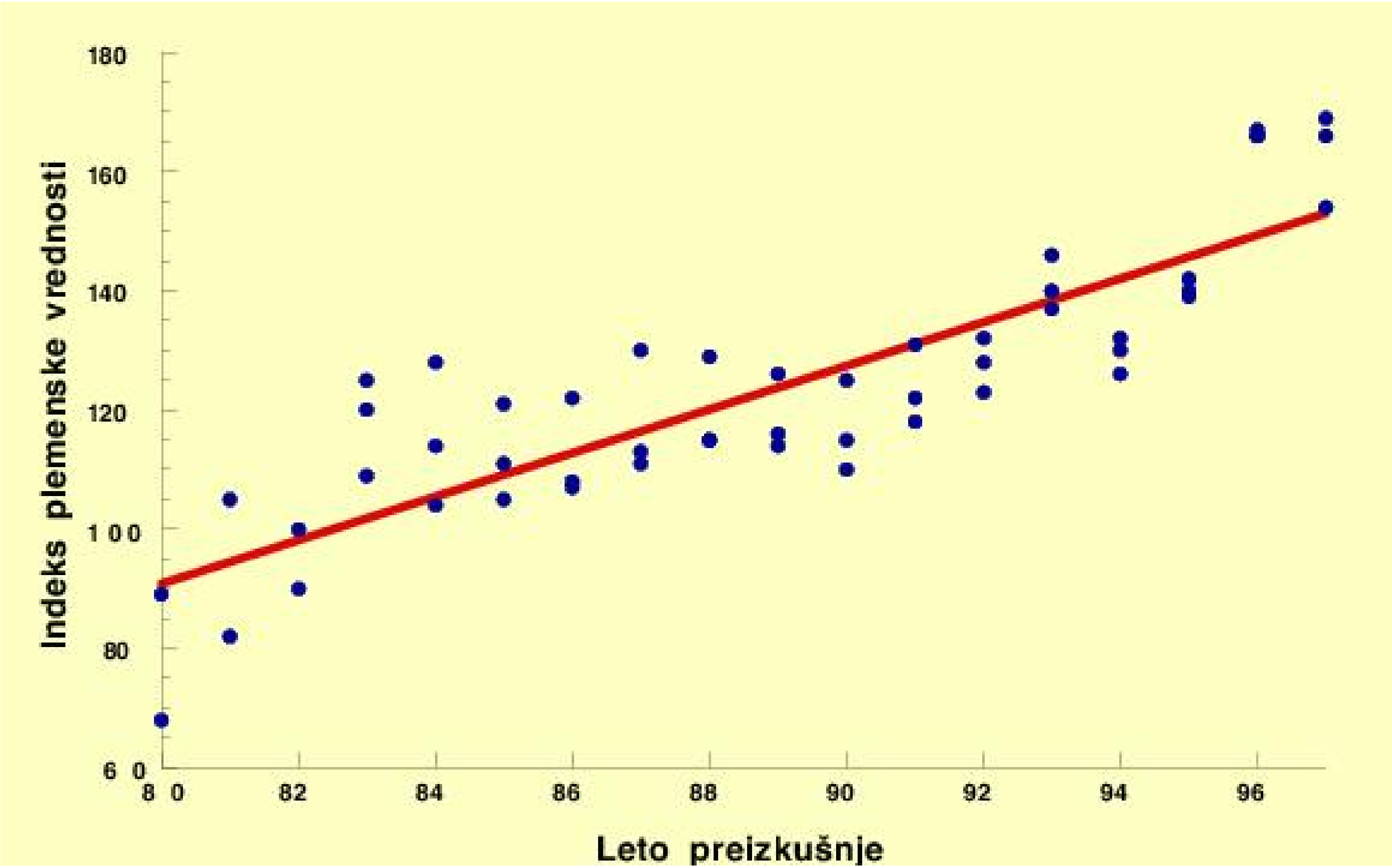
Regresijski koeficient: $\widehat{b} = 0$ točk / leto

Ekvivalentni model: $y_i = \mu + e_i$

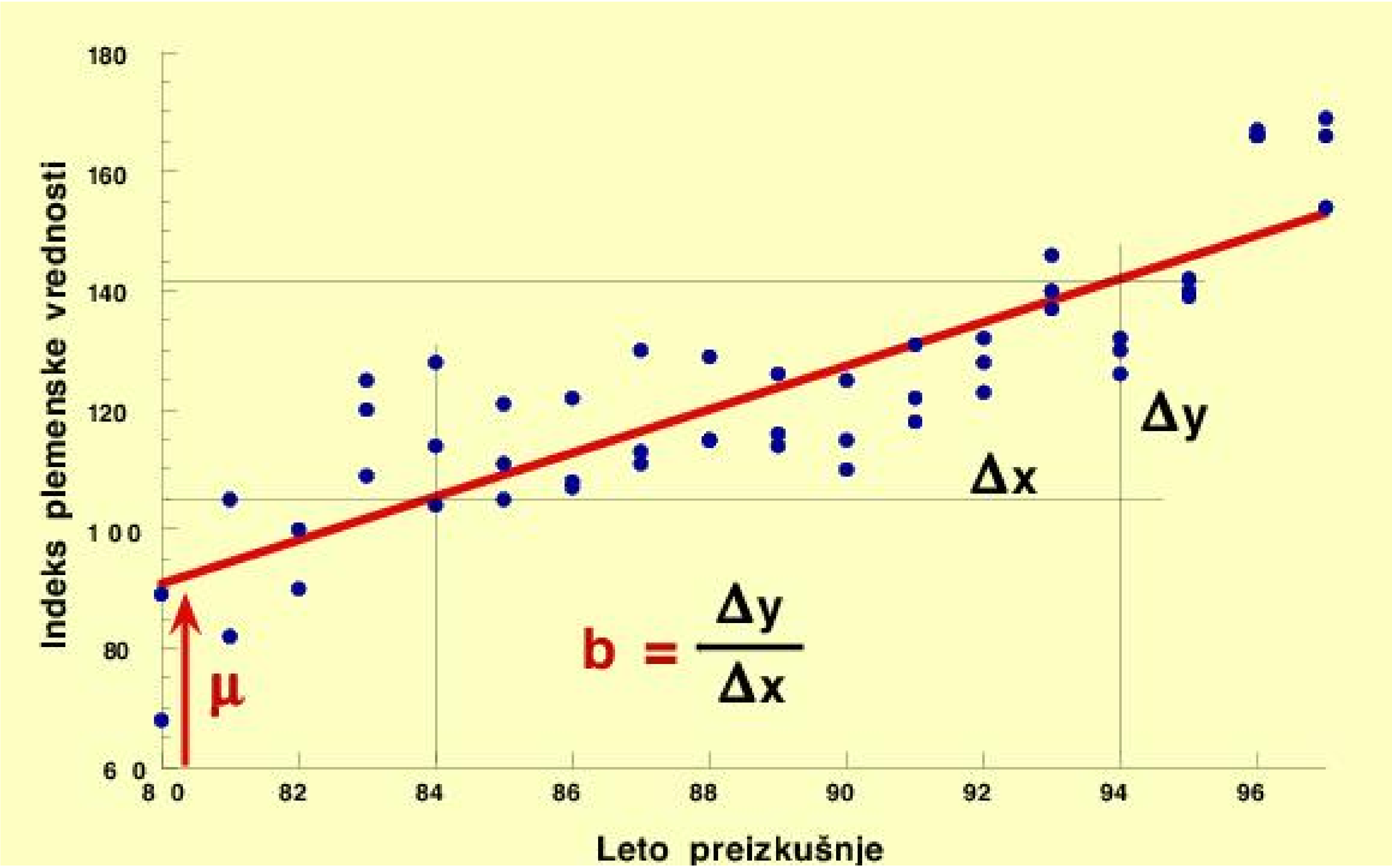
Vzporedno z x-osjo?



Linearna regresija



Linearna regresija



Odgovorite

Regresijski koeficient: $\frac{(142-105)}{(94-84)} \doteq 3.7$točke / leto

Model: $y_i = \mu + b\,(x_i - 80) + e_i$

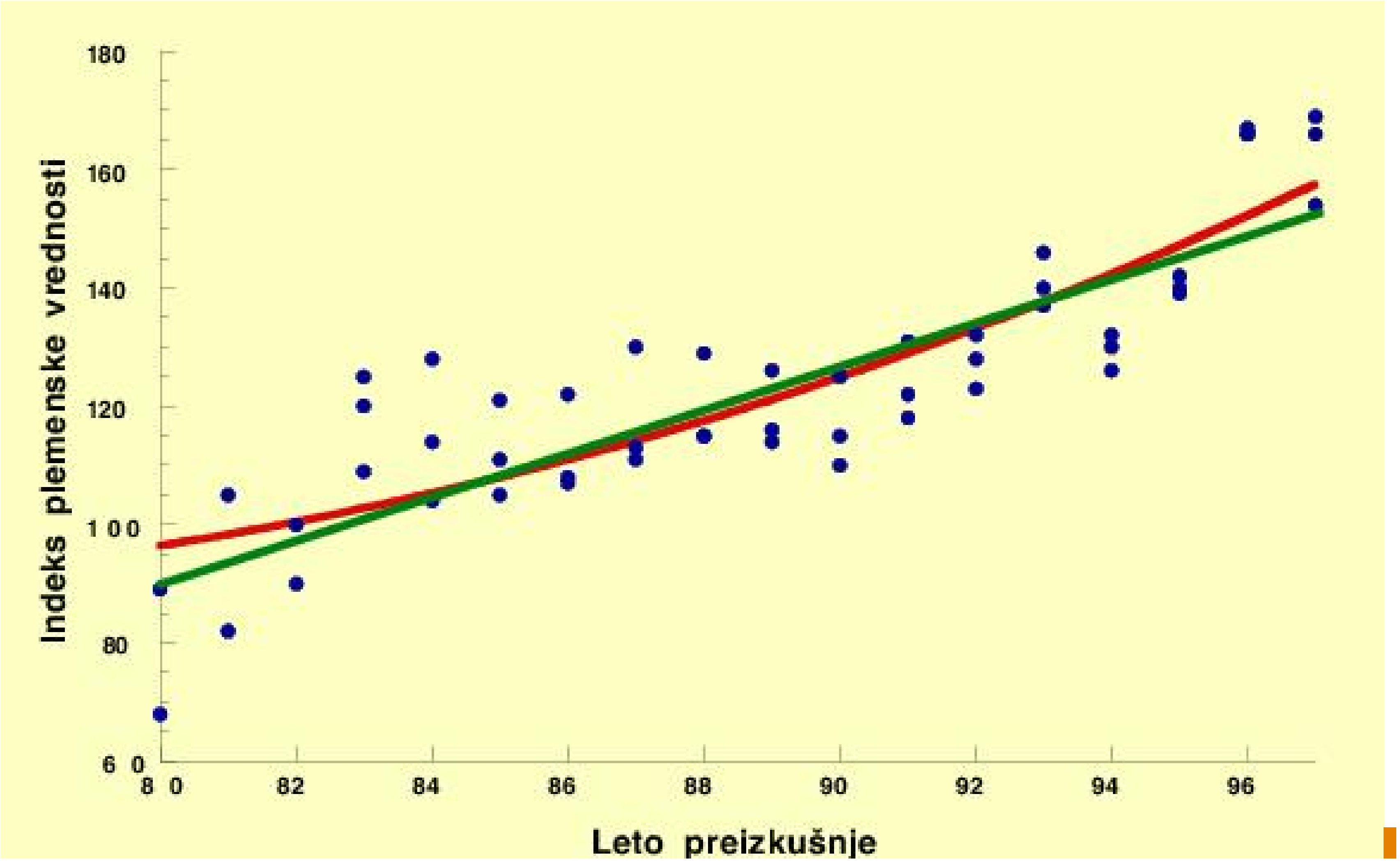
Obrazložitev povezave:

Indeks PV narašča za 3.7 točk na leto. V 10 letih se je tako indeks PV povečal za 37 točk.

Popravek? Ali smo povezavo dovolj dobro opisali?

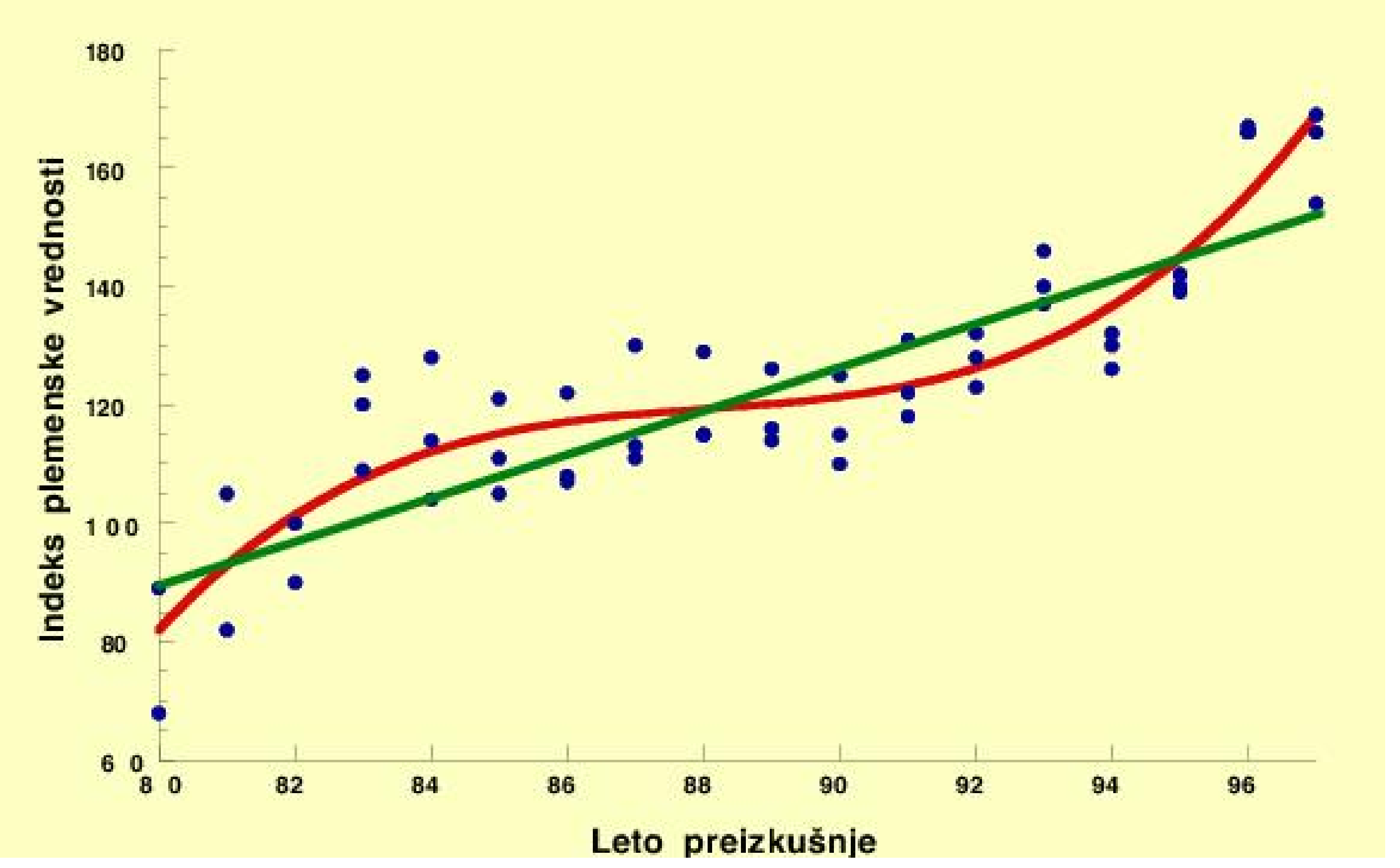
Povezava je dokaj dobro opisana, a lahko bi bilo bolje ...

Polinom druge stopnje



$$y_i = \mu + b_I(x_i - 80) + b_{II}(x_i - 80)^2 + e_i$$

Polinom tretje stopnje



Napišimo enačbo modela!

Polinom tretje stopnje

$$y_i = \mu + b_I (x_i - 80) + b_{II} (x_i - 80)^2 + b_{III} (x_i - 80)^3 + e_i$$

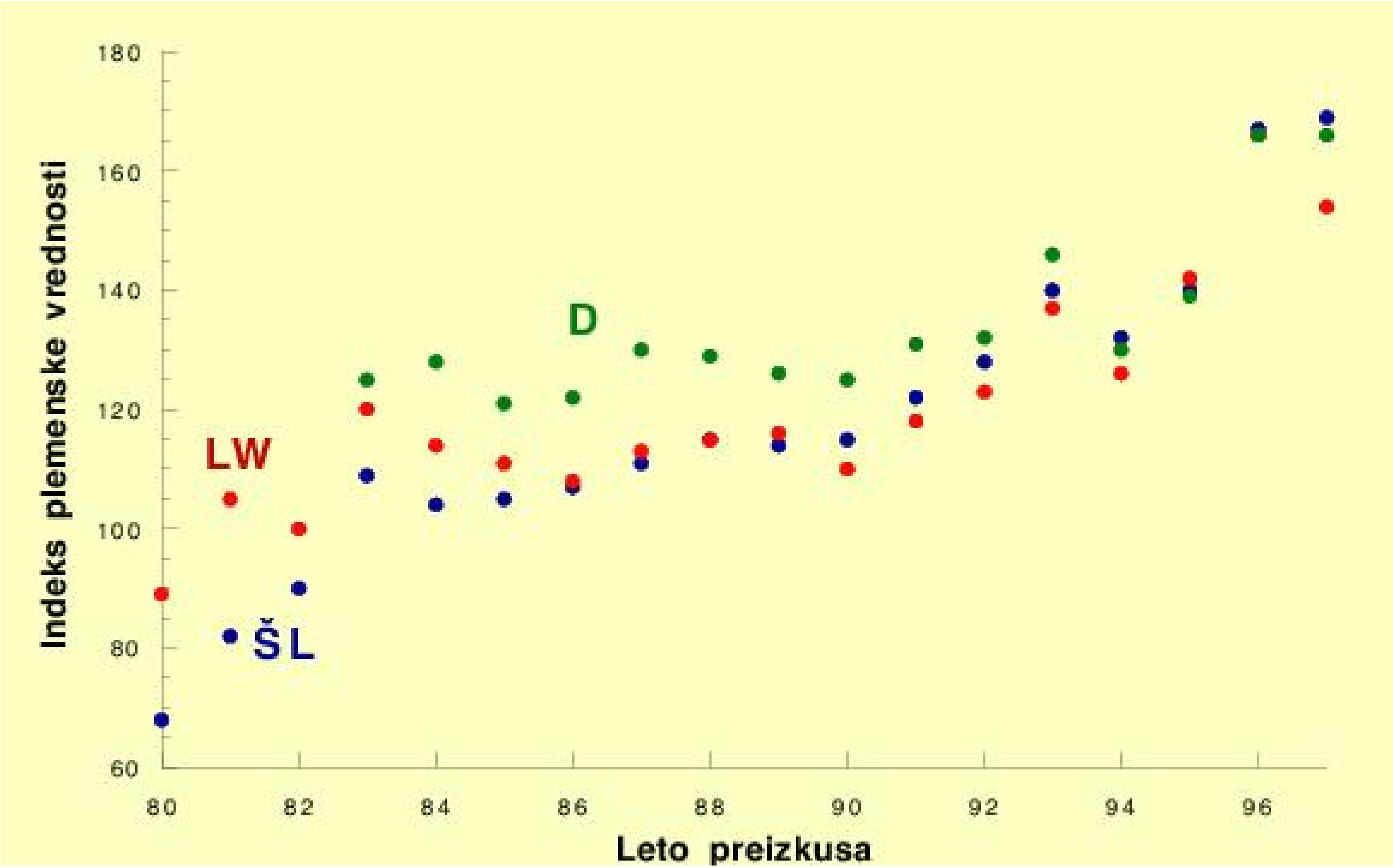
Linearni člen $\cdots + b_I (x_i - 80)^1 +$

Kvadratni člen $+ b_{II} (x_i - 80)^2 +$

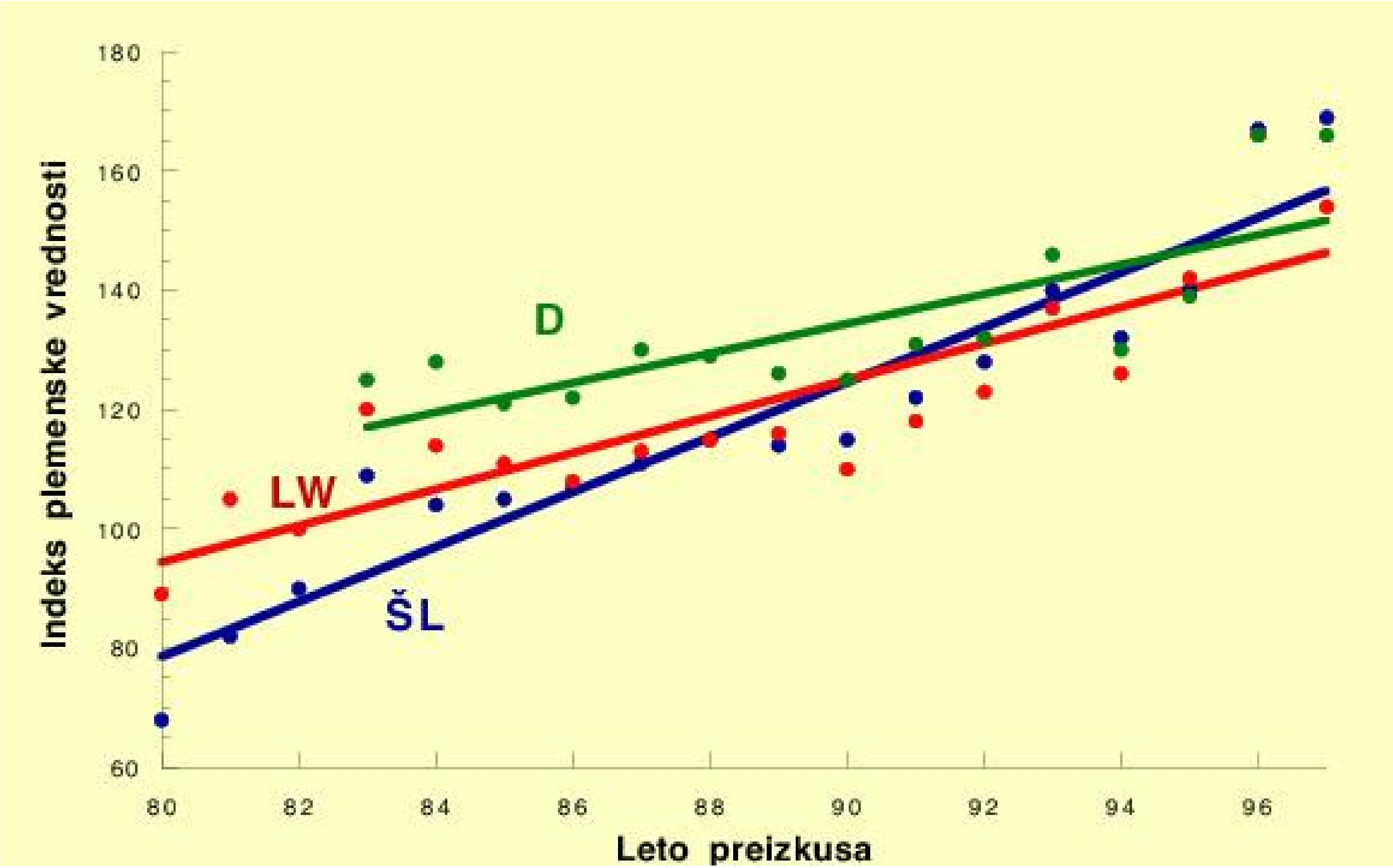
Kubni člen $+ b_{III} (x_i - 80)^3 + \cdots$

Zamolčana resnica o podatkih

Na grafu prikazujemo indeks PV pri treh pasmah!



Ugnezdene regresijske enačbe



Sestavimo model!

Neodvisne spremenljivke: pasma (kvalitativna s.)
leto preizkušnje (x, kvantitativna in diskretna s., regresija)

Odvisna spremenljivka: indeks plemenske vrednosti (y, kvantitativna in zvezna s., normalna porazdelitev)

Model: $y_{ij} = \mu + P_i + b_i(x_{ij} - 80) + e_{ij}$

Ekvivalentni model: $y_{ij} = \mu_i + b_i(x_{ij} - 80) + e_{ij}$

Pomnimo: regresija je ugnezdena znotraj vpliva pasme
vsaka pasma ima svojo regresijsko premico

Ugnezdene premice: ocene

$\widehat{\mu}_1 = 78.5$ točk $\widehat{b}_1 = 4.6$ točk / leto

$\widehat{\mu}_2 = 94.4$ točk $\widehat{b}_2 = 3.1$ točk / leto

$\widehat{\mu}_3 = 109.5$ točk $\widehat{b}_3 = 2.5$ točk / leto

Presečišča z y-osjo so izvrednotena za leto 1980!

Ekstrapolacija ni dovoljena!

... še nekaj ocen ...

$$y_{ij} = \mu_i + b_i(x_{ij} - 80) + e_{ij}$$

A			B			C		
P_1	x_{ij}	$\widehat{y}_{ij}$	P_2	x_{ij}	$\widehat{y}_{ij}$	P_3	x_{ij}	$\widehat{y}_{ij}$
1	80		2	80		3	80	
1	85		2	85		3	85	
1	90		2	90		3	90	
1	100		2	100		3	100	

(Letom bi morali prišteti 1900, vendar pa smo raje ostali pri manjših vrednostih!)

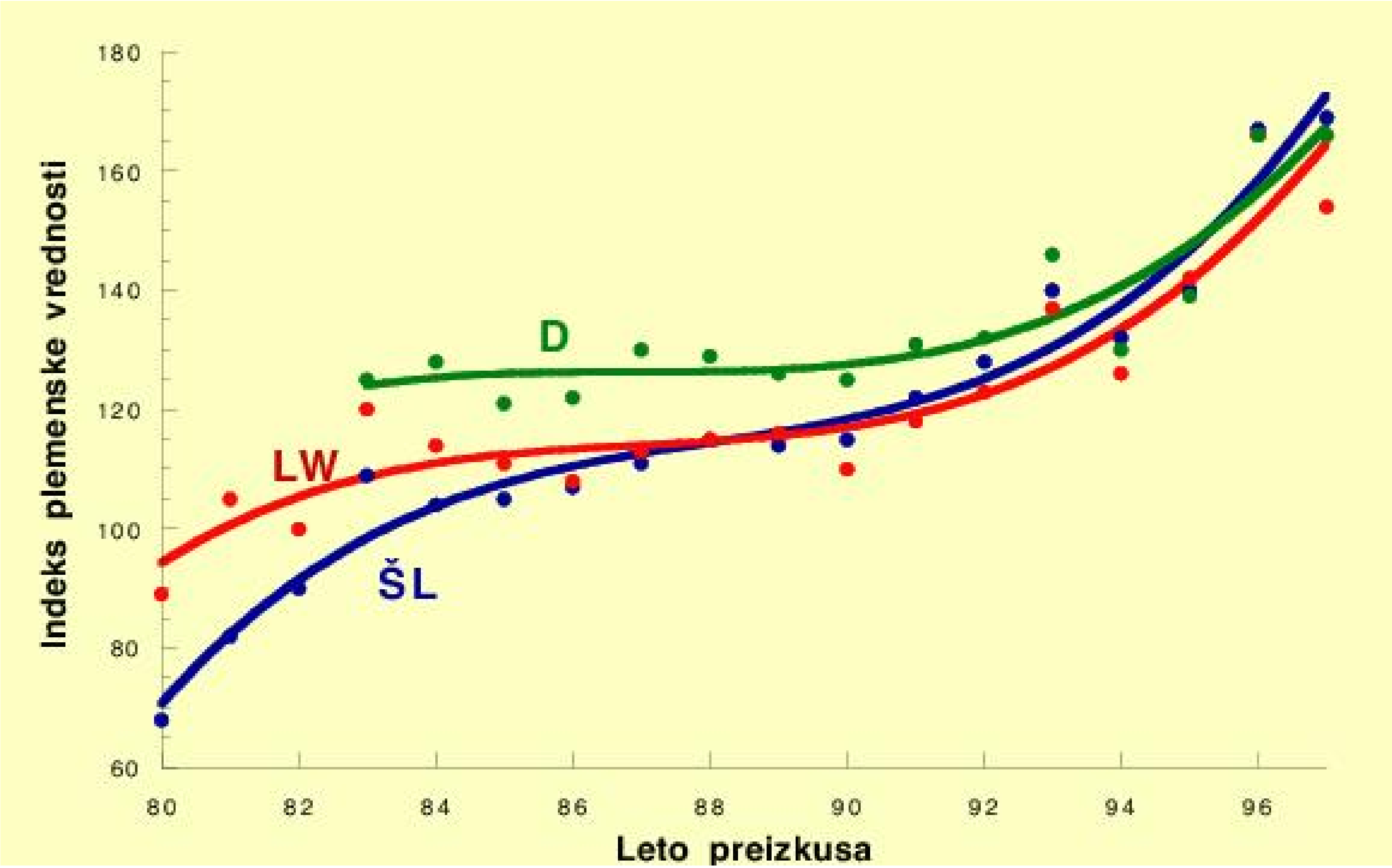
… še nekaj ocen …

$$\widehat{y}_{ij} = \widehat{\mu}_i + \widehat{b}_i (x_{ij} - 80)$$

A			B			C		
P	x_{ij}	$\widehat{y}_{ij}$	P	x_{ij}	$\widehat{y}_{ij}$	P	x_{ij}	$\widehat{y}_{ij}$
1	80	78.5	2	80	94.4	3	80	109.5
1	85	101.5	2	85	109.9	3	85	122.0
1	90	124.5	2	90	125.4	3	90	134.5
1	100	170.5	2	100	156.4	3	100	159.5

Ekstrapolacija ni dovoljena!

Ugnezdeni polinomi



Krivulje rišemo samo na intervalu, na katerem imamo podatke.

Ugnezdeni polinomi: ocene

- Možna sta dva ekvivalentna modela

$$y_{ij} = \mu + P_i + b_{Ii}(x_{ij} - 80) + b_{IIi}(x_{ij} - 80)^2 + b_{IIIi}(x_{ij} - 80)^3 + e_{ij}$$

$$y_{ij} = \mu_i + b_{Ii}(x_{ij} - 80) + b_{IIi}(x_{ij} - 80)^2 + b_{IIIi}(x_{ij} - 80)^3 + e_{ij}$$

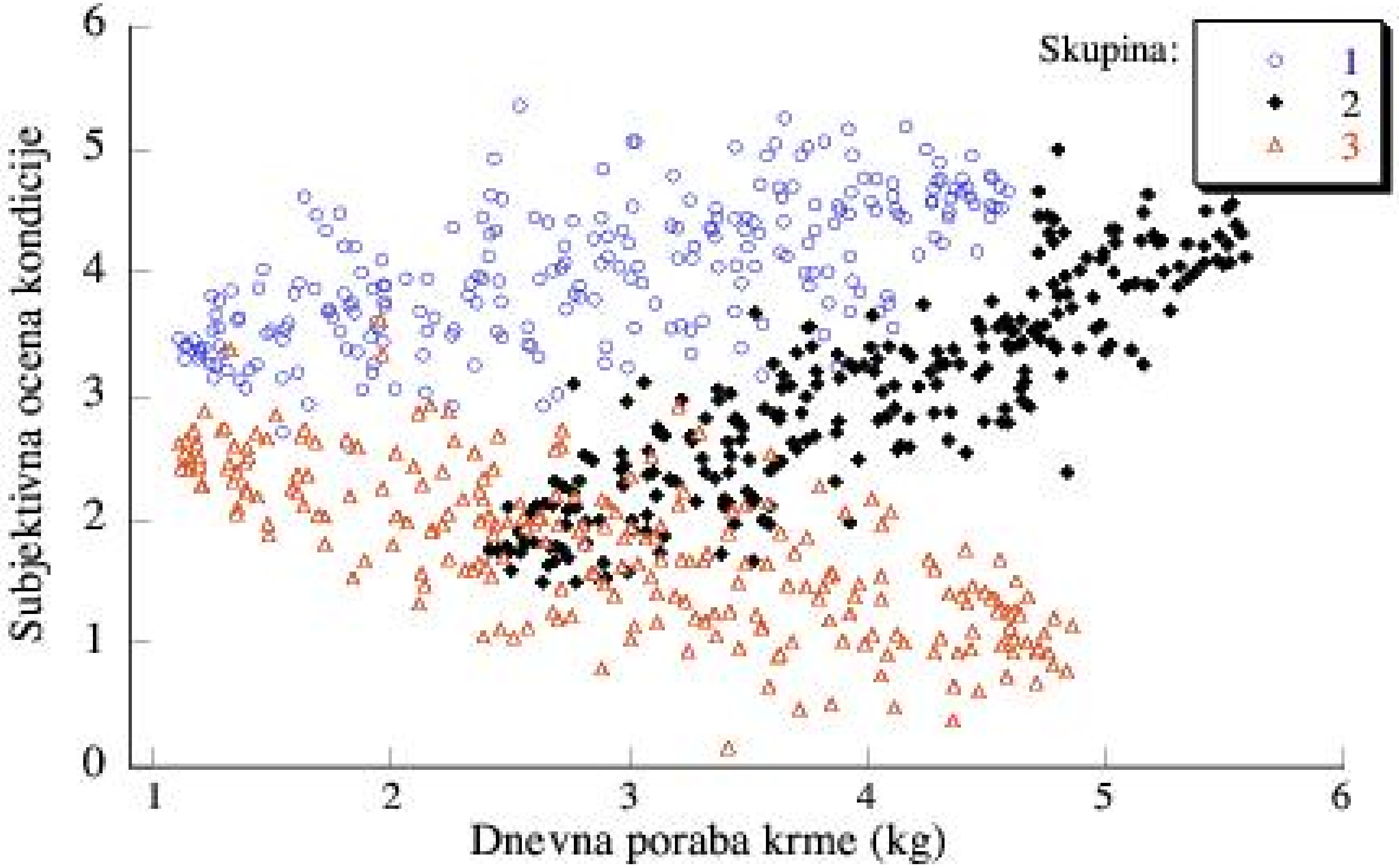
Enote t		$t/leto$		$t/leto^2$		$t/leto^3$
$\hat{\mu}_1 =$	$\widehat{b}_{I1} =$	1357.4	$\widehat{b}_{II1} =$	-15.41	$\widehat{b}_{III1} =$	0.0584
$\hat{\mu}_2 =$	$\widehat{b}_{I2} =$	1022.1	$\widehat{b}_{II2} =$	-11.74	$\widehat{b}_{III2} =$	0.0450
$\hat{\mu}_3 =$	$\widehat{b}_{I3} =$	888.0	$\widehat{b}_{II3} =$	-10.22	$\widehat{b}_{III3} =$	0.0392

- t –točk

Dogovor pri polinomih!

- regresijske koeficiente označujemo z malo črko b
- neodvisno spremenljivko označimo z x
- pri polin. višje stopnje dodamo indeks za stopnjo polin.
  - praviloma rimsko, izjemoma arabsko številko
- potenca pri neodvisni spremenljivki je enaka indeksu pri regresijskem koeficientu
- pri določanju stopnje polinoma pomagamo si lahko tudi z grafom

Dnevni poraba krme in subjektivna ocena



(podatki so simulirani in ne prikazujejo dejanske povezave!)

Sestavimo model!

Vplivi: skupina (kvalitativni v., z nivoji, S_i),
dnevna poraba krme (x, kvantit. vpliv, 3 premice)

Neodvisna sprem.: dnevna poraba krme (regresija)

Odvisna sprem.: subjektivna ocena (y, N-porazd.)

Model: $y_{ij} = \mu + S_i + b_i x_{ij} + e_{ij}$

Ekvivalentni model: $y_{ij} = \mu_i + b_i x_{ij} + e_{ij}$

Parametri

Regresijski koeficienti: $b_1,\ b_2,\ b_3$

Presečišča z y-osjo: $S_1,\ S_2,\ S_3$ ali $\mu_1,\ \mu_2,\ \mu_3$

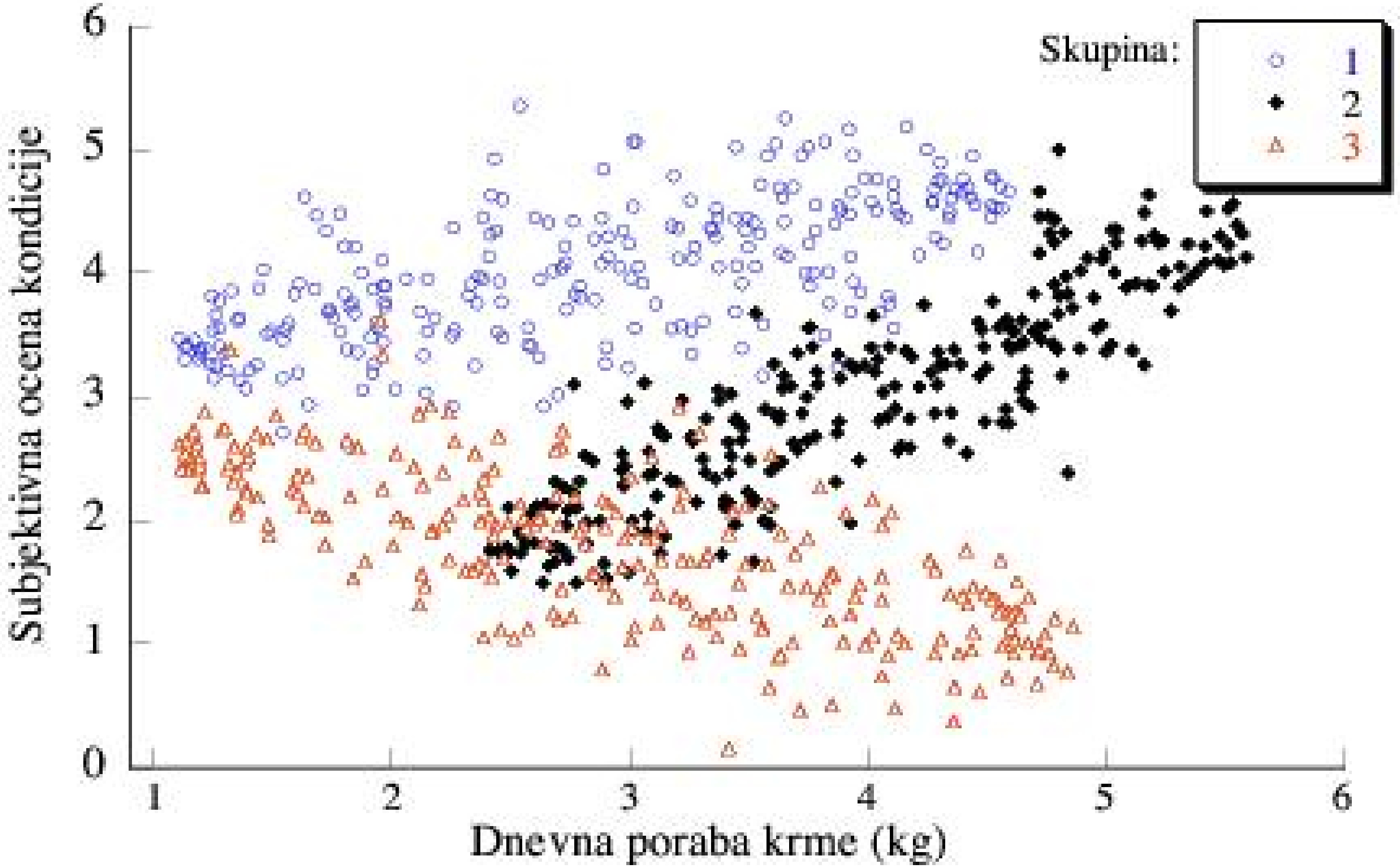
Parametri: $\mu,\ S_1,\ S_2,\ S_3,\ b_1,\ b_2,\ b_3$

Število parametrov: $1+3+3=7$

Število stopinj prostosti: $1+(3-1)+3=6$

Komentar:

Dnevni poraba krme in subjektivna ocena



Komentar

Subjektivne ocene (SO) so odvisne od dnevne porabe krme (DPK). Ko DPK narašča, se v modri in črni skupini tudi SO povečujejo. V črni skupini so vrednosti na opazovanem intervalu nižje kot v modri skupini, se pa povečujejo hitreje. SO v rdeči skupini pa z naraščanjem DPK padajo. ... dodamo diskusijo, primerjavo z literaturo, zakaj odstopanja ...

Ali smo povezavo dovolj dobro opisali? Da.

Subjektivne ocene

- subjektivne ocene imajo praviloma le nekaj ocen
  - skala 1 - 3, samo celo vrednosti
  - skala 1 - 10, samo cele vrednosti
  - skala 1 - 5, dovoljene tudi vmesne ocene (korak 0.5)
  - porazdelitev ni zvezna, zato ni normalna
  - če je razporeditev ocen simetrična in v skladu z normalno porazdelitvijo, lahko pri obdelavi predostavimo normalno porazdelitev
- kondicijo smo ocenjevali zvezno (izjema), podatke smo simulirali s pomočjo normalne porazdelitve, zato lahko privzamemo, da je porazdelitev normalna

Stopinje prostosti

- število ocenljivih parametrov

- koliko parametrov moram oceniti, da vem vse o vplivu?

  - parameter μ: ocenimo iz podatkov, 1 s.p.
  - parameter S_1: ocenimo iz podatkov, 1 s.p.
  - parameter S_2: ocenimo iz podatkov, 1 s.p.
  - parameter S_3: izračunamo, ne ocenjujemo iz podatkov, 0 s.p.

$$\mu = \frac{1}{3}(S_1 + S_2 + S_3) \Rightarrow S_3 = 3\mu - (S_1 + S_2)$$

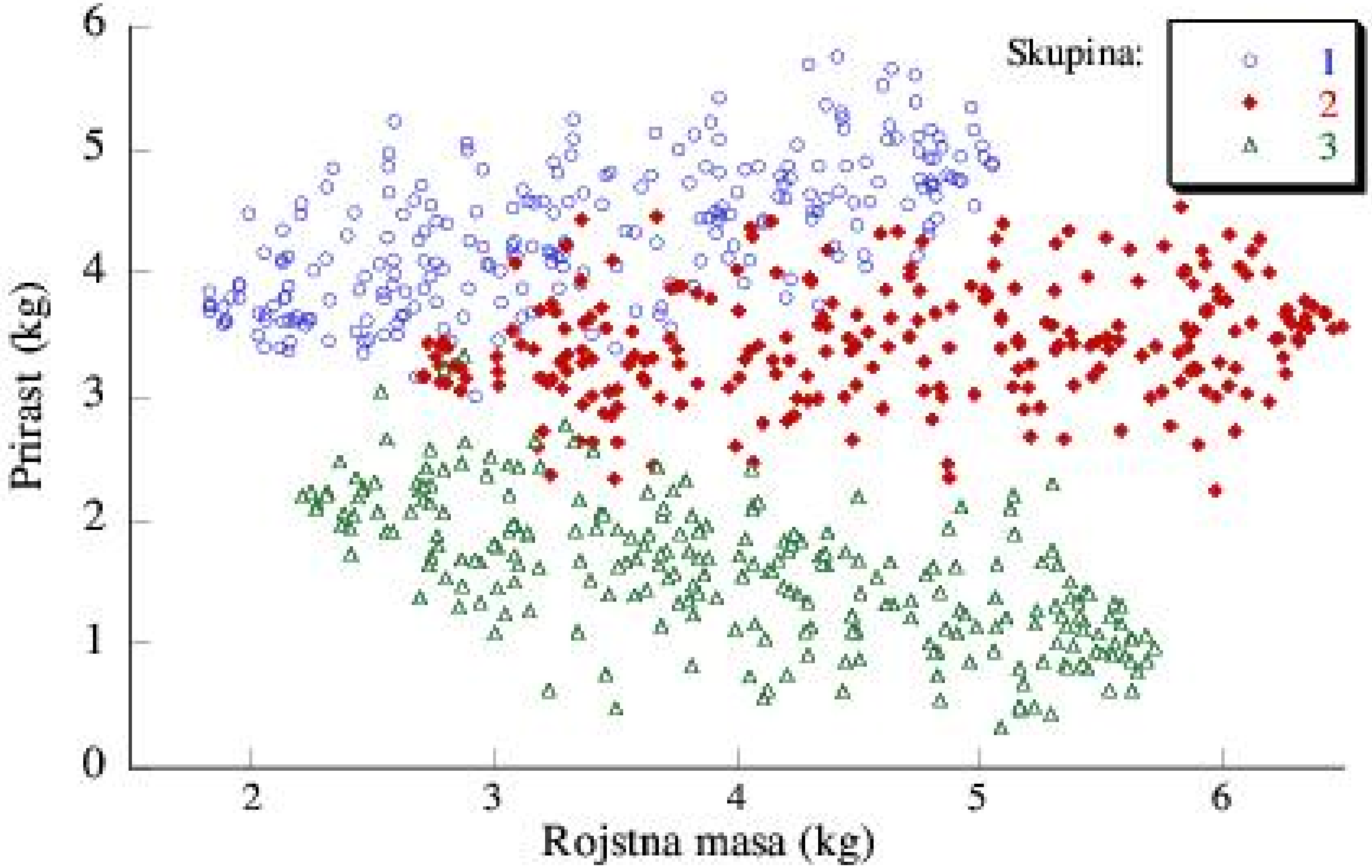
Stopinje prostosti

- regresija
  - parameter b_1: ocenimo iz podatkov, 1 s.p.
  - parameter b_2: ocenimo iz podatkov, 1 s.p.
  - parameter b_3: ocenimo iz podatkov, 1 s.p.
- oceniti moramo vsak regresijski koeficient, skupaj 3
- tako imamo 3 stopinje prostosti

Stopinje prostosti - povzetek

Parameter	Število param.	Število stopinj prostosti	Pojasnilo
μ	1	1	
Glavni vplivi			
z nivoji	p	$p-1$	$\mu = \frac{1}{p}\sum S_i$
Regresije	p	p	
Ugnezdeni vplivi	$\sum_i^p q_i$	$\sum (q_i - 1)$	$A_i = \sum_j^{q_i} B_{ij}$
Interakcije	$\sum pq$	$\sum (p-1)(q-1)$	
			$A_i = \sum_j^q AB_{ij}$

Rojstna masa in prirast



(podatki so simulirani in ne prikazujejo dejanske povezave!)

Sestavimo model!

Vplivi: skupina (kvalitativni v., z nivoji, S_i),
rojstna masa (x, kvantitativni vpliv)

Neodvisna spremenljivka: rojstna masa (3 premice)

Odvisna spremenljivka: prirast (y, N-porazdelitev)

Model: $y_{ij} = \mu + S_i + b_i x_{ij} + e_{ij}$

Ekvivalentni model: $y_{ij} = \mu_i + b_i x_{ij} + e_{ij}$

Parametri

Regresijski koeficienti: b_1, b_2, b_3

Presečišča z y-osjo: $S_1 = S_2 = S_3$ ali $\mu_1 = \mu_2 = \mu_3$

vse tri premice imajo isto presečišče z y-osjo

Sestavimo model! (nadalj.)

Primeren model: $y_{ij} = \mu + b_i x_{ij} + e_{ij}$

Komentar: ... poskusimo, čeprav so zaključki neuporabni ...
Primer prikazuje izjemo, ko smemo izpustiti vpliv, ki ima ugnezdeno regresijo ...

Ali smo povezavo dobro opisali? Da.

Določite število parametrov v vseh treh modelih!

Določite število stopinj prostosti!

Določite število opazovanj v vseh treh modelih!
(Število živali v skupinah je različno.)

Primerjava modelov

Število opazovanj v poskusu: $n = n_1 + n_2 + n_3$

$y_{ij} = \mu + S_i + b_i x_{ij} + e_{ij}$	μ	S_i	b_i	model	ostanek
Število parametrov	1	3	3	7	—
Število stopinj prostosti	1	2	3	6	$n-6$
$y_{ij} = \mu_i + b_i x_{ij} + e_{ij}$	μ_i		b_i	model	ostanek
Število parametrov	3		3	6	—
Število stopinj prostosti	3		3	6	$n-6$
$y_{ij} = \mu + b_i x_{ij} + e_{ij}$	μ		b_i	model	ostanek
Število parametrov	1		3	4	—
Število stopinj prostosti	1		3	4	$n-4$

Ekvivalentni modeli

Modeli so ekvivalentni:

- če imajo isto število ocenljivih parametrov
- če ocenljivi parametri zagotavljajo iste zaključke
- opisujejo podatke precej podobno (isti vplivi, samo druga oblika)

Prva dva modela na zadnji tabeli sta ekvivalentna:

- v prvem modelu parameterS_3 ni ocenljiv,
- v drugem nismo vključili skupne srednje vrednosti (μ), ampak srednje vrednosti za posamezne skupine (μ_i)

Debelina hrbtne slanine

✔ DHS se pri prašičih z maso med 90 in 110 kg povečuje za nekako 0.10 mm/kg. Če vemo, da se je masa spremenila za 10 kg, pričakujemo lahko za 1 mm debelejšo DHS. Pri 100 kg je npr. povprečna DHS 13 mm.

1. Odvisna spremenljivka:

2. Neodvisna spremenljivka:

3. Regresijski koeficient:

4. Narišimo graf!

Debelina hrbtne slanine

✔ **DHS** se pri prašičih z **maso** med 90 in 110 kg povečuje za nekako 0.10 mm/kg. Če vemo, da se je masa spremenila za 10 kg, pričakujemo lahko za 1 mm debelejšo DHS. Pri 100 kg je npr. povprečna DHS 13 mm.

1. Odvisna spremenljivka: **Debelina hrbtne slanine**
2. Neodvisna spremenljivka: **maso**
3. Regresijski koeficient: 0.10 mm/kg
4. Narišimo graf!

Ocene debeline hrbtne slanine

- Napišimo enačbo za oceno DHS od 90 do 100 kg!

$$\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 (x_i - 100) = 13.0 + 0.10 * (x_i - 100)$$

- Parametra b_0 in b_1 sta ocenjeni z ocenama $\widehat{b}_0$ in $\widehat{b}_1$

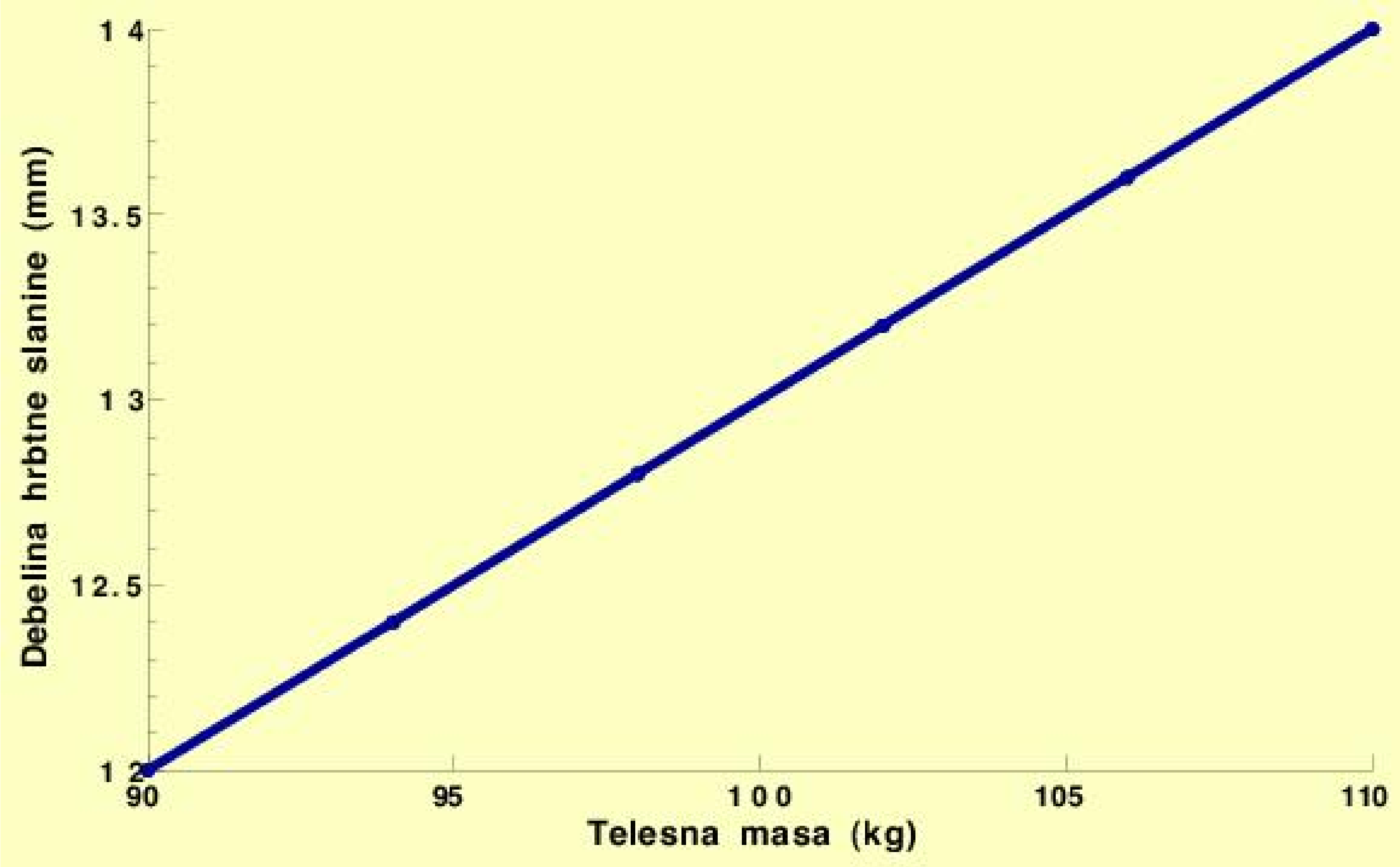
- Izračunajmo nekaj točk (za premico najmanj dve!)

Masa (kg)	90	94	98	100	102	106	110
DHS (mm)	12.0	12.4	12.8	13.0	13.2	13.6	14.0

Kadar se mudi, zadostujeta za premico samo dve točki!

- Sedaj pa še narišimo graf!

Graf za debelino hrbtne slanine



Pomni!

1. Regresijski koeficient ima vedno sestavljeno enoto: v števcu od odvisne spremenljivke, v imenovalcu pa od neodvisne spremenljivke (npr. mm/kg, točk/leto)

2. Za ponazoritev uporabimo GRAF S ČRTAMI. Neodvisno spremenljivko nanesemo na os X, odvisno na os Y.

3. Kvantitativne vplive praviloma opišemo z regresijo. Izjeme so izredno redke.

4. Porazdelitev neodvisne sprem. prilagodimo poskusu!

5. Pri regresiji s pomočjo neodvisne (pojasnjevalne) spremenljivke ocenimo, pojasnimo odvisno spremenljivko

Načrt poskusa pri kvantitativnih vplivih

1. Neodvisna spremenljivka - določimo interval

2. Opazovanja enakomerno pozazdelimo na intervalu

3. Več opazovanj naredimo lahko
   (a) pri ekstremnih vrednosti neodvisne spremenljivke
   (b) kjer pričakujemo zavoje

4. Neodvisna spremenljivka ima lahko
   (a) samo nekatere vrednosti (npr. 90, 94, 98, 100 kg...)
   (b) vse vrednosti na intervalu
   (c) porazdelitev NI POMEMBNA!