Statistični modeli

Milena Kovač

16. november 2012

Statistični modeli

1. Izvleček iz opazovanj

2. Abstrakcija realnosti

3. Poenostavljena slika, prirejena iz podatkov

4. Opazovanje pojasnimo z vplivi

5. Nepojasnjeni del razlik ostane v ostanku

$$lastnost = funkcija\,[vplivi] + ostanek$$

Sinonimi za lastnost

- $y_{ij\ldots}$, samo izjemoma kaj drugega
- meritve, subjektivne ocene/točke
- opazovanja
- odvisne spremenljivke
- “posledica”

Vplivi

- lastnosti so odvisne od številnih vplivov
- "vzroki"
- vplivi so lahko znani (zabeleženi)
- lahko pa so tudi neznani (spregledani)
- lahko so pomembni (značilni)
- ali nepomembni (zanemarljivi, niso značilni)

Ostanek

- Napaka (naključna napaka, error)
  - napaka pri meritvah (samo do sprejemljive tolerance!)
  - napaka pri izvedbi poskusa
- Ostanek (zavestna napaka, residual)
  - pozabljeni in zanemarjeni vplivi
    (neznanje ni opravičilo za površnost!)
  - poenostavljen model
    (izpustimo lahko le vplive z majhnim učinkom)

Elementi statističnega modela

1. Enačba modela (equation): $y_{ij..} =?$

2. Pričakovane vrednosti (expected values): $E(y_{ij..}) =?$

3. Struktura varianc in kovarianc (covariance structure, covariance matrices): $cov(y_{ij..}) =?$

4. Predpostavke (assumptions) in omejitve (restrictions): opišemo v stavkih

Primer (ovce)

✔ Pri ovcah dveh pasem so proučevali maso jagnjet. Jagnjeta so bila odstavljena hkrati, vendar so bila zaradi različnega datuma rojstva različno stara...

Opazovanje	masa ob odstavitvi	y
Vplivi	pasma	P
	starost	x
Regresijski	koeficient	b
Ostanek		e

Primer (ovce)

Pasma	i	Jagnje	j	Odstavitev Starost (dni)	Masa (kg)	Masa ob prodaji (kg)
Texel		1		90	38.7	40.2
Texel		2		85	35.3	37.0
Texel		3		95	32.1	36.7
JS ovca		4		88	26.2	32.5
JS ovca		5		91	27.3	28.9
JS ovca		6		97	33.4	35.6
P_i		(a_{ij})		x_{ij}	y_{ij}	

Primer (ovce)

Pasma	i	Jagnje	j	Odstavitev Starost (dni)		Odstavitev Masa (kg)		Masa ob prodaji (kg)
Texel	1	1	1	90	x_{11}	38.7	y_{11}	40.2
Texel	1	2	2	85	x_{12}	35.3	y_{12}	37.0
Texel	1	3	3	95	x_{13}	32.1	y_{13}	36.7
JS ovca	2	4	1	88	x_{21}	26.2	y_{21}	32.5
JS ovca	2	5	2	91	x_{22}	27.3	y_{22}	28.9
JS ovca	2	6	3	97	x_{23}	33.4	y_{23}	35.6
P_i		(a_{ij})		x_{ij}		y_{ij}		

Primer (ovce)

✔ Pri ovcah dveh pasem so proučevali maso jagnjet. Jagnjeta so bila odstavljena hkrati, a so bila … različno stara …

Opazovanje	Masa ob odstavitvi	y_{ij}	
Srednja vrednost		μ	
Vplivi	Pasma	P_i	$i=1,\ 2$
	Starost	x_{ij}	
Regresijski	koeficient	b	
Ostanek		e_{ij}	$j=1,\ 2,\ 3$

Enačba modela:

$$y_{ij} = \mu + P_i + b x_{ij} + e_{ij}$$

Enačba

$$y_{ij} = \mu + \ldots + e_{ij}$$

Oznake v modelu

Opazovanje: y

Srednja vrednost: μ ali b_0 ali α

Ostanek: e

Oznake v modelu: vplivi

$$\ldots + P_i + bx_{ij} + \ldots$$

1. Sistematski vplivi:

   (a) vpliv z nivoji: P
   ena črka, **velika**, brez strešic, spominja na ime vpliva
   tudi grške črke: μ , β ...

   (b) regresijski koeficient: b ali β

   (c) neodvisna spremenljivka: x

2. Naključni vplivi:

   (a) ena črka, **mala**, brez strešic, spominja na ime vpliva (a_{ij})

Indeksi v modelu

$$y_{ij} = \mu + P_i + bx_{ij} + e_{ij}$$

- $i = 1,\ 2 \Leftarrow$ker sta dve pasmi
- $j = 1,\ 2,\ \ldots\ n_i \Leftarrow$ označuje opazovanja znotraj pasme
- $ij \Leftarrow$ določa vsako opazovanje
- opazovanje ima vedno iste indekse kot napaka,
- pogosto tudi neodvisna spremenljivka x:
  pri vsaki meritvi naredimo napako
  (tudi $e_{ij} = 0$ je napaka!)

Parametri (ovce)

$$y_{ij} = \mu + P_i + bx_{ij} + e_{ij}$$

- opazovanje: odstavitvena masa jagnjet y_{ij}
- ostanek: e_{ij}
- vplivi: pasma (P), starost (jagnjeta j pri pasmi i - x_{ij})
- parametri (neznanke v modelu): μ, P_1, P_2, b

Parametri (nadalj.)

- parametri (neznanke v modelu): μ, P_1, P_2, b

- število (neznanih) parametrov: $1+2+1=4$

- število opravljenih meritev: $n = n_1 + n_2 = \sum_i n_i$

Ne zamenjujte vplivov, parametrov in ocen

Vplivi: vzroki za razlike (razpršenost)

- pasma (P) in starost (x_{ij})

Parametri: neznana vrednost za posamezne nivoje

- masa jagnjet pri texel (P_1) in JS ovci (P_2)
- povečevanje odstavitvene mase s starostjo (b)

Ocena: iz podatkov izračunana vrednost parametra

- jagneta tehtajo $\widehat{P_1}$ pri texel in ($\widehat{P_2}$) JS pasmi
- masa jagnet s starostjo narašča z $\widehat{b}$ kg/dan

Primer II - ovce

- dodajmo modelu še vpliv živali (naključni vpliv, kvalitativni),

- starost pa opišimo s kvadratno regresijo ugnezdeno znotraj pasme

$$y = \mu + P + b\left(x - \begin{Bmatrix} 0 \\ \overline{X} \\ c \end{Bmatrix}\right) + b\left(x - \begin{Bmatrix} 0 \\ \overline{X} \\ c \end{Bmatrix}\right)^2 + a + e$$

Primer II - regresijski koeficienti

- imamo dva različna regresijski koeficienta

  - za linearni člen b_I
  - za kvadratni člen b_{II}

$$y = \mu + P + b_I \left(x - \left\{ \begin{array}{c} 0 \\ \overline{X} \\ c \end{array} \right\} \right) + b_{II} \left(x - \left\{ \begin{array}{c} 0 \\ \overline{X} \\ c \end{array} \right\} \right)^2 + a + e$$

Primer II - kvalitativni vplivi

- dodajmo indekse kvalitativnim vplivom (vplivom z razredi)

  - imamo dve pasmi P_i
  - pri vsaki pasmi P_i imamo več živali a_{ij}
  - žival a_{ij} pripada pasmi P_i

$$y = \mu + P_i + b_I \left(x - \left\{ \begin{array}{c} 0 \\ \overline{X} \\ c \end{array} \right\} \right) + b_{II} \left(x - \left\{ \begin{array}{c} 0 \\ \overline{X} \\ c \end{array} \right\} \right)^2 + a_{ij} + e$$

Primer II - lastnosti in ostanki

- dodajmo indekse za lastnosti in ostanke

  - žival a_{ij} ima samo eno meritev y_{ij} in en ostanek e_{ij}
  - če bi žival imela več meritev, bi jih šteli z dodanim indeksom (npr. y_{ijk})

$$y_{ij} = \mu + P_i + b_I \left(x - \left\{ \begin{array}{c} 0 \\ \overline{X} \\ c \end{array} \right\} \right) + b_{II} \left(x - \left\{ \begin{array}{c} 0 \\ \overline{X} \\ c \end{array} \right\} \right)^2 + a_{ij} + e_{ij}$$

Primer II - kvantitativni vplivi

- dodajmo indekse za neodvisno spremenljivko starost
  - žival a_{ij} ima eno meritev y_{ij} pri starosti x_{ij}
  - pomeni, da pasmi rasteta enako hitro

$$y_{ij} =$$

$$\mu + P_i + b_I \left(x_{ij} - \left\{ \begin{array}{c} 0 \\ \overline{X} \\ c \end{array} \right\} \right) + b_{II} \left(x_{ij} - \left\{ \begin{array}{c} 0 \\ \overline{X} \\ c \end{array} \right\} \right)^2 + a_{ij} + e_{ij}$$

Primer II - starost ugnezdena znotraj pasme

- pomeni, da pasmi rasteta lahko različno hitro

- za vsako pasmo P_i rabimo par regresijskih koeficientov:
  - za linearni člen b_{Ii}
  - za kvadratni člen b_{IIi}

$$y_{ij} =$$

$$\mu + P_i + b_{Ii}\left(x_{ij} - \left\{\begin{array}{c} 0 \\ \overline{X} \\ c \end{array}\right\}\right) + b_{IIi}\left(x_{ij} - \left\{\begin{array}{c} 0 \\ \overline{X} \\ c \end{array}\right\}\right)^2 + a_{ij} + e_{ij}$$

Primer II - konstante v modelu

$$y_{ij} = \mu + P_i + b_{Ii}\left(x_{ij} - \left\{\begin{array}{c} 0 \\ \overline{X} \\ c \end{array}\right\}\right) + b_{IIi}\left(x_{ij} - \left\{\begin{array}{c} 0 \\ \overline{X} \\ c \end{array}\right\}\right)^2 + a_{ij} + e_{ij}$$

- presečišče y z x osjo
- ocene (rešitve) so korigirane na skupno vrednost 0, povprečje $\overline{X}$ ali konstanto c

$$\widehat{y}_{ij} = \widehat{\mu} + \widehat{P}_i + 0 + 0 \text{ ; če } x_{ij} = \left\{\begin{array}{c} 0 \\ \overline{X} \\ c \end{array}\right\}$$

- med modeli ni pomembne razlike - so ekvivalentni

Primer - študenti biometrije

✔ Pročiti želimo vpliv sedežnega reda na znanje biometrije. Klopi so razvrščene v vrste in stolpce.

Vplivi: vrsta - V_i, i=1, 2, 3, 4, 5

Vplivi: stolpec - S_j, j=1, 2, 3, 4, 5

Enačba modela:

$$y_{ijk} = \mu + V_i + S_j + e_{ijk}$$

- v klopi sedi več študentov: k=1, 2, ... n_{ij}

Parametri - študenti biometrije

$$y_{ijk} = \mu + V_i + S_j + e_{ijk}$$

- merili smo: y_{ijk}
- ostanek: e_{ijk}
- vplivi: vrsta (V), stolpec (S)
- parametri: μ, V_1, V_2, V_3, V_4, V_5, V_6, V_7, S_1, S_2, S_3
- število parametrov: $1 + 7 + 3 = 11$

Število študentov poskusu

$$y_{ijk} = \mu + V_i + S_j + e_{ijk}$$

- $ij \Rightarrow$ označuje klop v vrsti i in stolpcu j
- $k \Rightarrow$ označuje študente po klopeh ij
- $n = \sum n_{ij} \Leftarrow$vsota števila ($n_{ij}$) študentov po klopeh
- enako število študentov po klopeh $\Rightarrow n = m_i * m_j * m_k$
  pomnožimo število vrst (m_i) s številom klopi v vrsti (m_j) in številom študentov v klopi (m_k)
- če v vsaki klopi sedijo 4 študenti$\Rightarrow n = 7 * 3 * 4 = 81$

Ali je klop pomembna?

$$y_{ijk} = \mu + V_i + S_j + VS_{ij} + e_{ijk}$$

- oznaka VS_{ij} predstavlja klop v vrsti V_i in stolpcu S_j
- uporabimo jo, če bi bilo kakorkoli možno, da bi klop ugodno ali neugodno vplivala na rezultate študentov v njej (popisana klop, lahka dostopnost pedagoga ...)
- poseben, specifičen vpliv, dodatno poleg splošnega trenda
- tak sestavljen vpliv je interakcija
- oznake dobi od sestavljajočih vplivov
- indeksi po abecednem vrstnem redu

Poskusimo I

- Vplivi A, R in C so sistematski, kvalititativni in križno klasificirani (prepleteni)

- Vpliv x je sistematski, kvantitativni, npr. polinom druge stopnje ...

- Vplivi g, a in p so naključni, kvalititativni, ugnezdeni znotraj vseh prej navedenih kvalitativnih vplivov

$$\begin{aligned} y \quad = \quad & \mu + C \quad + R \quad + A \quad + \\ & +b \quad (x \quad - \qquad) + b \quad (x \quad - \qquad)^2 + \\ & +g \quad + a \quad + p \quad + e \end{aligned}$$

- Vpišite v model še dvojne in trojne interakcije med vplivi A, R in C

- Ena meritev za vsak razred/nivo pri vplivu p

Preverimo rezultat I

- Vplivi A, R in C so sistematski, kvalititativni in križno klasificirani
- Vpliv x je sistematski, kvantitativni ...
- Vplivi g, a in p so naključni, kvalititativni, ugnezdeni znotraj vseh prej navedenih kvalitativnih vplivov
- Vpišite v model še dvojne in trojne interakcije med vplivi A, R in C
- Ena meritev za vsak razred/nivo pri vplivu p

$$
\begin{aligned}
y_{ijklmn} \;=\;& \mu + C_i + R_j + A_k + CR_{ij} + CA_{ik} + RA_{jk} + CRA_{ijk} + \\
& + b_I(x_{ijklmno} - \overline{x}) + b_{II}(x_{ijklmno} - \overline{x})^2 + \\
& + g_{ijkl} + a_{ijklm} + p_{ijklmn} + e_{ijklmn}
\end{aligned}
$$

Poskusimo II

- Vplivi C in R sta sistematska, kvalititativna in križno klasificirana
- A je sistematski, kvalititativni in ugnezden znotraj R
- Vpliv x_1 in x_2 sta sistematska, kvantitativna ...
- Vplivi g, a in p so naključni, kvalititativni, , ugnezdeni znotraj vpliva A, a znotraj g in p znotraj a

$$
\begin{aligned}
y \quad = \quad & \mu + C \quad + R \quad + A \quad + \\
& + b \quad (x_1 \quad - \qquad) + b \quad (x_2 \quad - \qquad) + \\
& + g \quad + a \quad + p \quad + e
\end{aligned}
$$

- Vpišite v model še dvojne in trojne interakcije med vplivi A, R in C

Preverimo rezultat II

- Vplivi C in R sta sistematska, kvalititativna in križno klasificirana
- A je sistematski, kvalititativni in ugnezden znotraj R
- Vpliv x_1 in x_2 sta sistematska, kvantitativna ...
- Vplivi g, a in p so naključni, kvalititativni, ugnezdeni znotraj vpliva A, a znotraj g in p znotraj a
- Vpišite v model še dvojne in trojne interakcije med vplivi A, R in C

$$
\begin{aligned}
y_{ijklmno} = {} & \mu + C_i + R_j + A_{jk} + CR_{ij} + \\
& + b_1(x_{1ijklmno} - \bar{x}_1) + b_2(x_{2ijklmno} - \bar{x}_2) + \\
& + g_{jkl} + a_{jklm} + p_{iklmn} + e_{ijklmno}
\end{aligned}
$$

Členi na desni strani modela

1. Regresijski koeficienti

2. Pojasnevalne spremenljivke

   (a) neodvisna spremenljivka
      i. kvantitativna spremenljivka
      ii. merimo, štejemo, točkujemo (ocenjujemo)
      iii. razvrstimo od manjše do večje vrednosti

   (b) vplivi z nivoji
      i. kvalitativna spremenljivka
      ii. nimajo vrednosti, razvrstitve so poljubne

Regresijski koeficient

1. Oznaka: b, b_i (mala črka "b") ali β, β_i (grška črka "β")

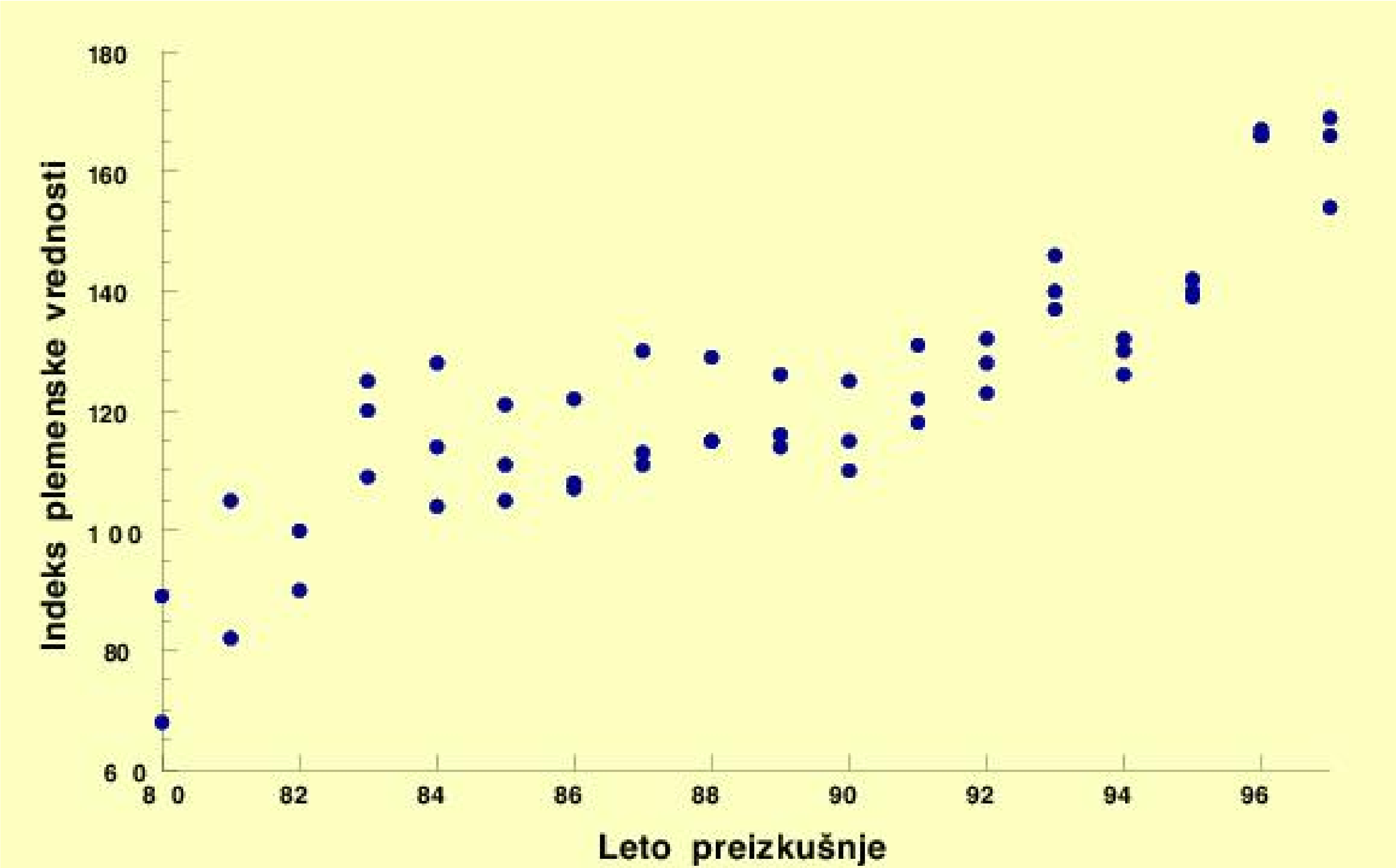
2. Smerni koeficient premice pri linearni regresiji

3. Sestavlja celoto z neodvisno spremenljivko x_i

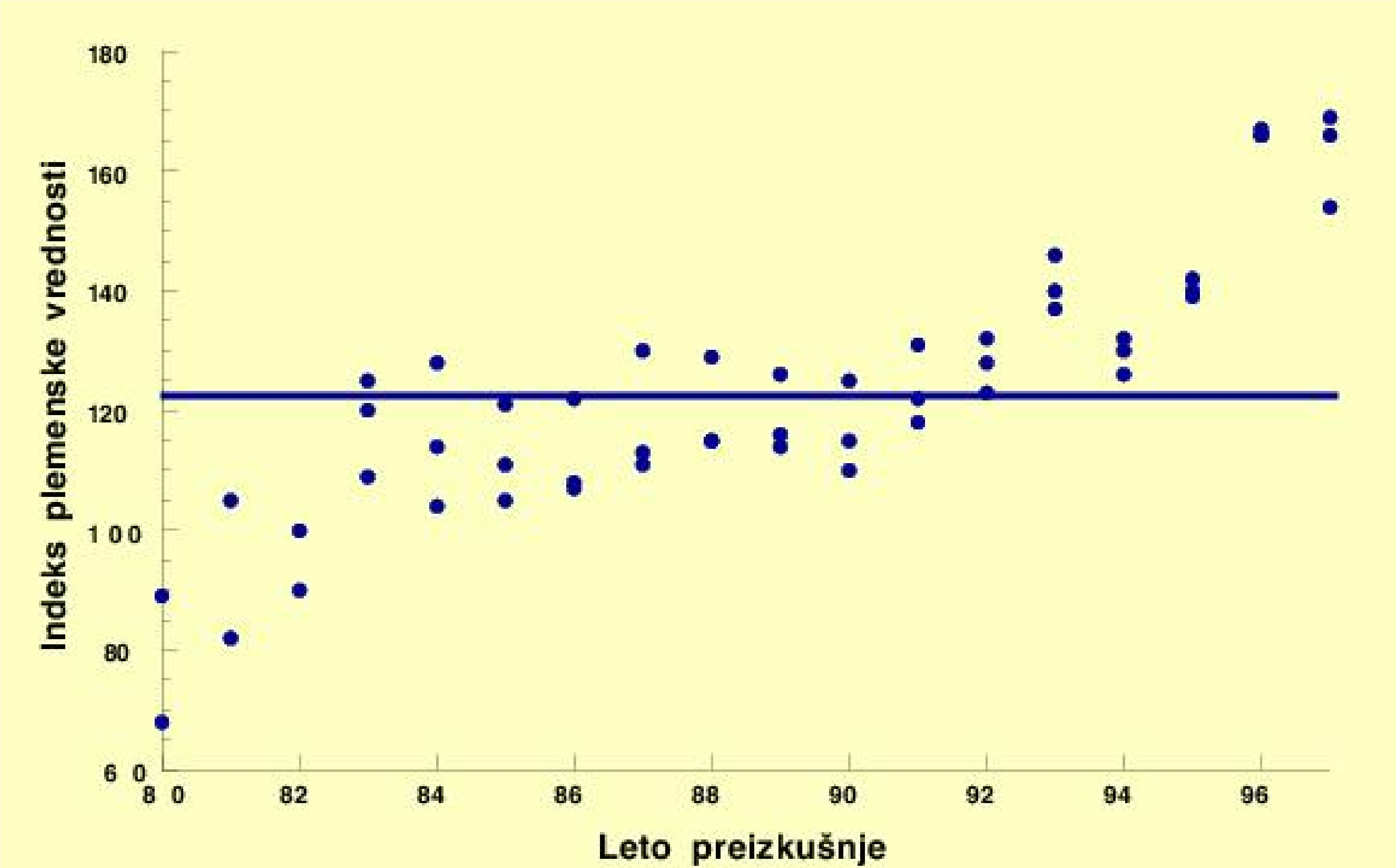
4. Primer linearne regresije

$$y_i = ... + \beta_1 * x_i + ...$$

Linearna regresija



Vzporedno z x-osjo?



Sestavimo model!

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

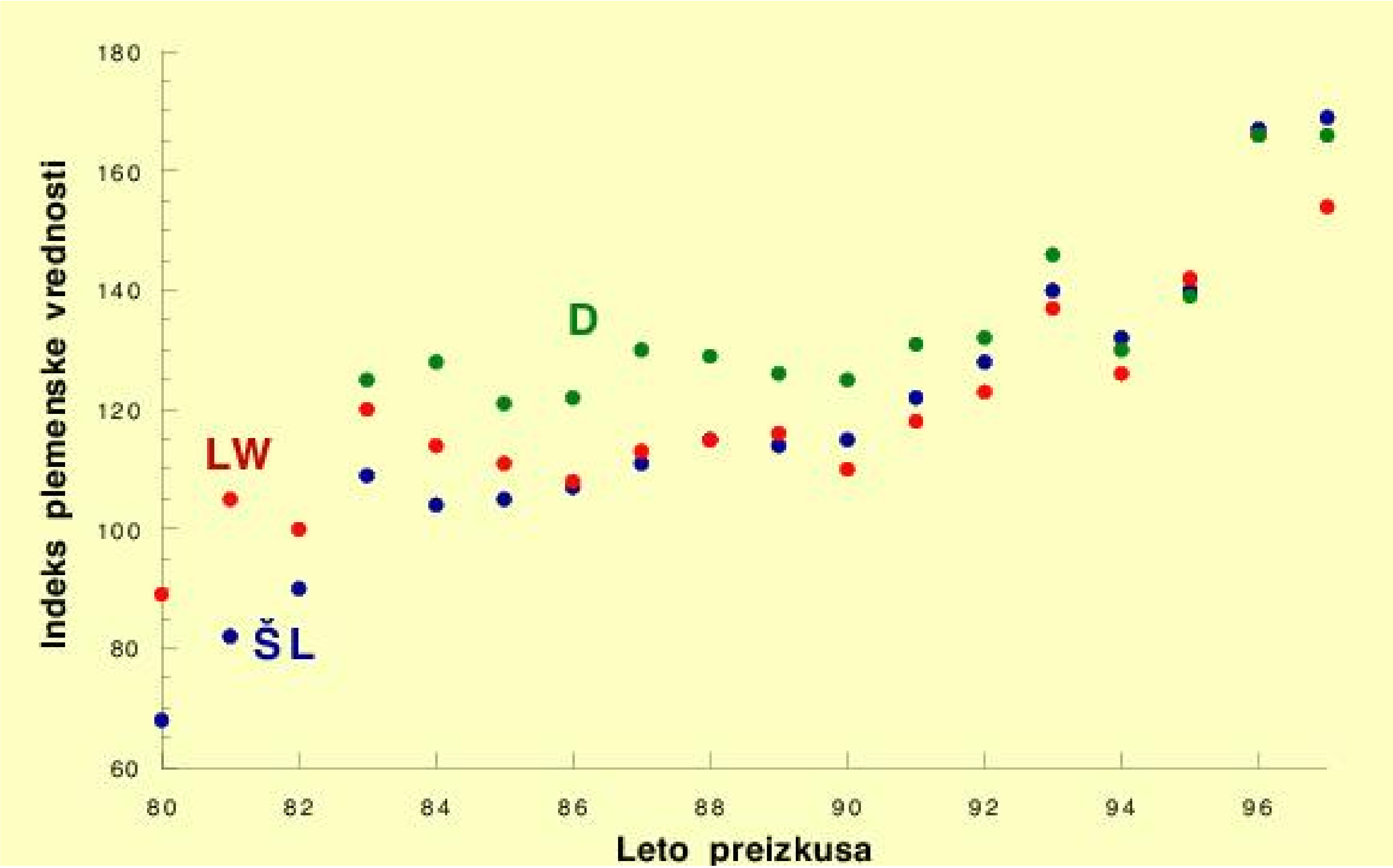
Odvisna spremenljivka: indeks plemenske vrednosti (y, kvantitativna in zvezna spremenljivka, N-porazdelitev)

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

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

Ekvivalentni model: $y_i = \mu + e_i$

Indeks plemenske vrednosti po pasmah



Model in indeksi - IPV

a) Napišimo model - vplivi z nivoji: pasma

$$y = \mu + P_i +$$

- P_i - sistematski vpliv pasme, $i = 1, 2, 3$;

b) Dodajmo kvantitativne vplive - katera funkcija je primerna?

$$y = \mu + P_i + b_I(x-80) + b_{II}(x-80)^2 + b_{III}(x-80)^3 +$$

c) Regresija je ugnezdena znotraj pasme:
b-ji dobijo indeks od pasme

$$y = \mu + P_i + b_{Ii}(x-80) + b_{IIi}(x-80)^2 + b_{IIIi}(x-80)^3 +$$

Model in indeksi - IPV

d) Dodajmo še ostanek:

$$y = \mu + P_i + b_{Ii}(x - 80) + b_{IIi}(x - 80)^2 + b_{IIIi}(x - 80)^3 + e$$

e) Opremimo odvisno spremenljivko in ostanek z indeksi:

- imamo več opazovanj $(j = 1, 2, \ldots n_i)$ za i-to pasmo
- vsaka meritev ima napako

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

f) Uredimo še neodvisno spremenljivko z indeksi:

- vsako leto (x_{ij}) imamo za i-to pasmo eno vrednost y_{ij}

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

Meritve debeline hrbtne mišice

Žival	Gnezdo	Pasma	Masa (kg)	D H M (mm) 1	2	3	Temp. peke (^0C)
1	1	11	113	81.3	79.4	80.2	180
2	1	11	106	76.0	74.0	75.2	80
3	2	11	119	81.0	83.4	85.0	180
4	2	11	107	93.2	92.0	92.8	80
5	3	22	101	74.9	73.6	72.7	180
6	3	22	106	84.8	84.2	85.6	80
7	4	22	117	75.8	78.0	76.3	180
8	4	22	120	88.2	84.9	86.5	80

Izpeljimo (sestavimo, nastavimo, napišimo) statistični model!

Glavni vplivi

Vpliv	sist./naklj.	kvant./kvalit.	oznaka	opomba
Žival	naključni	kvalitativni	a_{ijk}	znotraj gnezda
Gnezdo	naključni	kvalitativni	g_{ij}	znotraj pasme
Pasma	sistematski	kvalitativni	P_i	
Masa	sistematski	kvantitativni	$x_{m?}$	
Temperatura	sistematski	kvantitativni	$x_{t?}$	

Model in indeksi - DHM

a) Napišimo model - vplivi z nivoji: pasma, gnezdo, žival

$$y = \mu + P_i + g_{ij} + a_{ijk}$$

- P_i - sistematski vpliv pasme, $i = 1, 2, 3$;
- g_{ij} - naključni vpliv gnezda znotraj pasme, $j = 1, 2, \ldots n_i$;
- a_{ijk} - naključni vpliv živali znotraj pasme, $k = 1, 2, \ldots n_{ij}$

b) Dodajmo kvantitativne vplive

- sistematski vpliv - umestimo pred naključni del, brez indeksov
- določimo tudi konstanto

$$y = \mu + P_i + b_m(x_m - 110) + b_t(x_t - 80) + g_{ij} + a_{ijk}$$

Model in indeksi - DHM

c) Samo če je regresija ugnezdena znotraj pasme:

b-ji dobijo indeks od pasme (npr. temperatura)

presodimo po grafu, znanju, statistični preveritvi ...

$$y = \mu + P_i + b_m(x_m - 110) + b_{ti}(x_t - 80) + g_{ij} + a_{ijk}$$

d) Dodajmo še ostanek:

$$y = \mu + P_i + b_m(x_m - 110) + b_{ti}(x_t - 80) + g_{ij} + a_{ijk} + e$$

Model in indeksi - DHM

e) Opremimo odvisno spremenljivko (lastnost) in ostanek z indeksi:

žival a_{ijk} ima več ponovitev in sicer $l = 1, 2, \ldots n_{ijk}$

n_{ijk} je število ponovitev za vsako žival (v preglednici $n_{ijk} = 3$)

če so vsi $n_{ijk} = 1$, ne dodamo novega indeksa

$$y_{ijkl} = \mu + P_i + b_m(x_m - 110) + b_{ti}(x_t - 80) + g_{ij} + a_{ijk} + e_{ijkl}$$

f) Uredimo še neodvisno spremenljivko z indeksi:

vse tri meritve y_{ijkl} imajo za par isto vrednost x_{mijk} in x_{tijk}

pogosto so indeksi isti …

$$y_{ijkl} = \mu + P_i + b_m(x_{mijk} - 110) + b_{ti}(x_{tijk} - 80) + g_{ij} + a_{ijk} + e_{ijkl}$$

Alternativne možnosti modela za DHM

- namesto črk uporabimo (arabske) številke

$$y_{ijkl} = \mu + P_i + b_1(x_{1ijk} - 110) + b_{2i}(x_{2ijk} - 80) + g_{ij} + a_{ijk} + e_{ijkl}$$

- namesto konstante je lahko povprečje

$$y_{ijkl} = \mu + P_i + b_1(x_{1ijk} - \bar{x}_1) + b_{2i}(x_{2ijk} - \bar{x}_2) + g_{ij} + a_{ijk} + e_{ijkl}$$

- konstanta je lahko enaka tudi 0

$$y_{ijkl} = \mu + P_i + b_1 x_{1ijk} + b_{2i} x_{2ijk} + g_{ij} + a_{ijk} + e_{ijkl}$$

- namesto malih uporabimo velike črke (manj možnosti napak)

$$y_{ijkl} = \mu + P_i + b_M(x_{Mijk} - \bar{x}_M) + b_{Ti}(x_{Tijk} - \bar{x}_T) + g_{ij} + a_{ijk} + e_{ijkl}$$