Univerza v Ljubljani
Fakulteta za farmacijo

Multipla regresija

Iztok Grabnar

Univerza v Ljubljani, Fakulteta za farmacijo

Učenje/potrjevanje

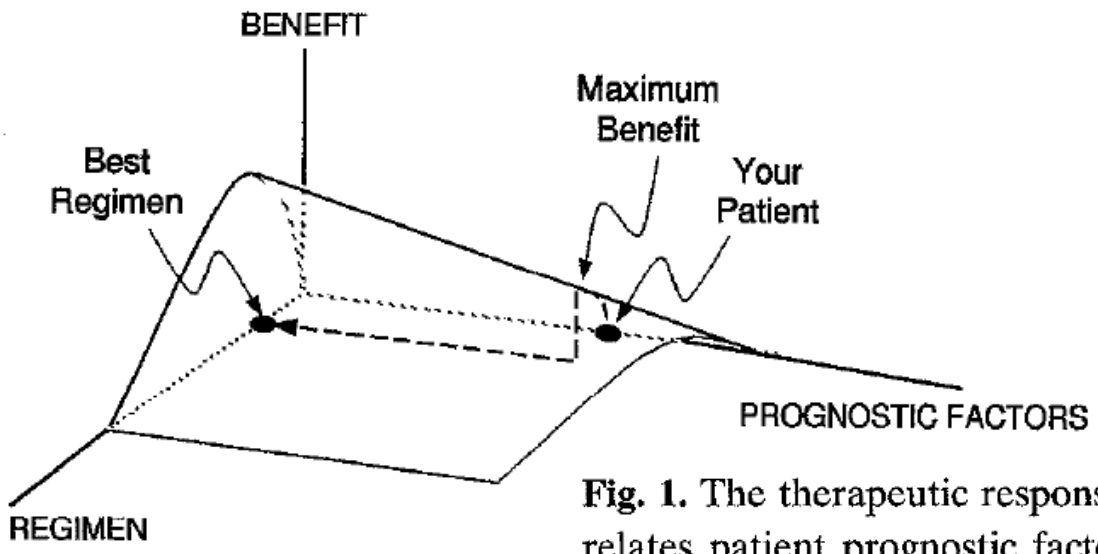
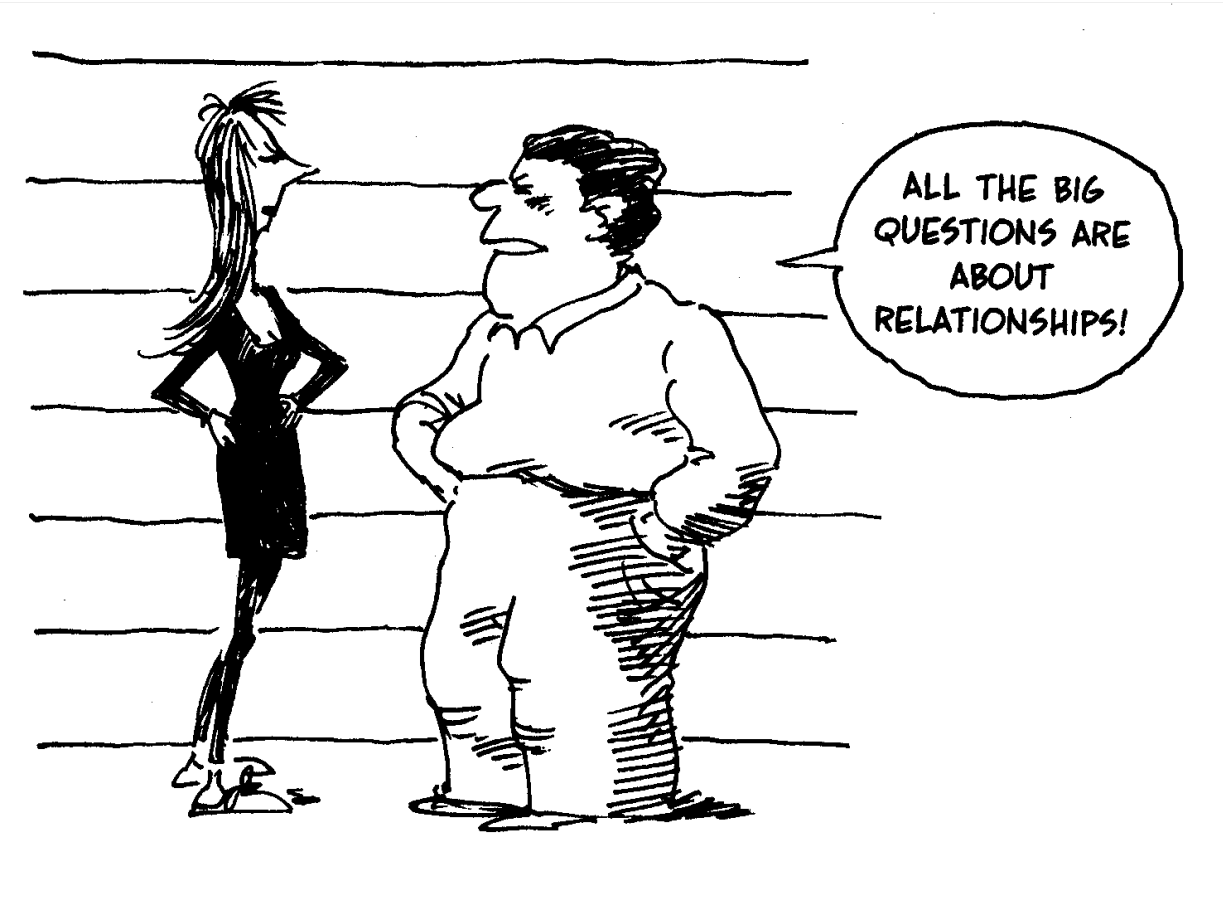


Fig. 1. The therapeutic response surface for a given drug relates patient prognostic factors (such as sex, age, and weight) and dose regimen (amount and timing) to benefit, the net utility of efficacy, and toxicity. A plane perpendicular to the prognostic axis at the value of "your" particular patient intersects the surface forming a curve (as shown). The optimal regimen for your patient is that corresponding to the maximum of this curve on the benefit axis (here, the value on the maximum benefit "ridge"). (Adapted from Sheiner. Clin Pharmacol Ther 1991;50:4-9. Used with permission.)


ALL THE BIG QUESTIONS ARE ABOUT RELATIONSHIPS!

Analiza povezanosti

- Opazovani pojav= odvisna spremenljivka
- Napovedni dejavnik= neodvisna spremenljivka
- Statistični modeli:
  - Univariabilni:
    - en napovedni dejavnik
    - Povezava kot pomembna pokaže:
      - zaradi dejanske povezanosti napovednega dejavnika s pojavom
      - lahko tudi zaradi povezanosti z nekim drugim napovednim dejavnikom.
  - Multivariabilni:
    - več napovednih dejavnikov

Regresijska analiza

- Univariatna, bivariatna, multivariatna analiza
- Statistična spremenljivka
  - Numerična zvezna
  - Atributivna
  - Dihotomna
- Linearna regresija
- Logistična regresija

Posplošeni linearni model
Generalized Linear Model

$$E(Y) = \mu = g^{-1}(X\beta)$$

Linearni prediktor

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

Povezovalna funkcija

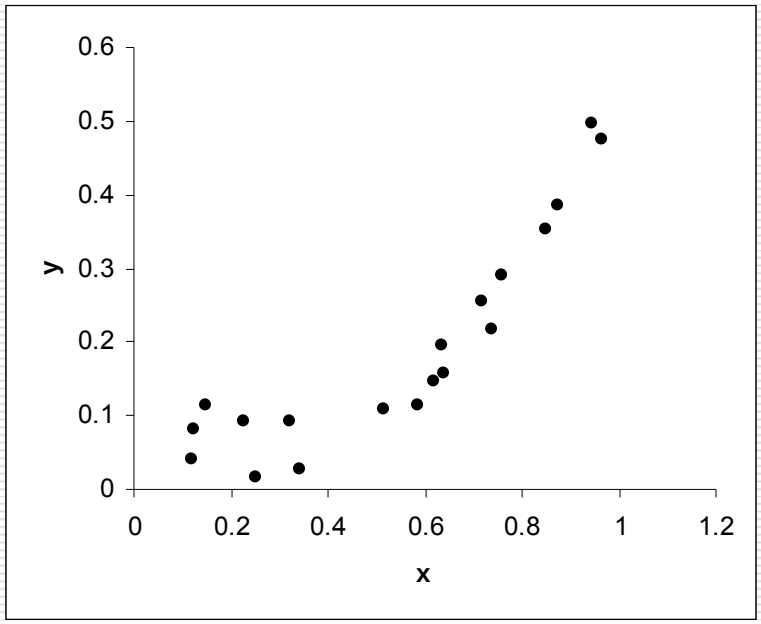
$$Var(Y) = V(\mu) = V(g^{-1}(X\beta))$$

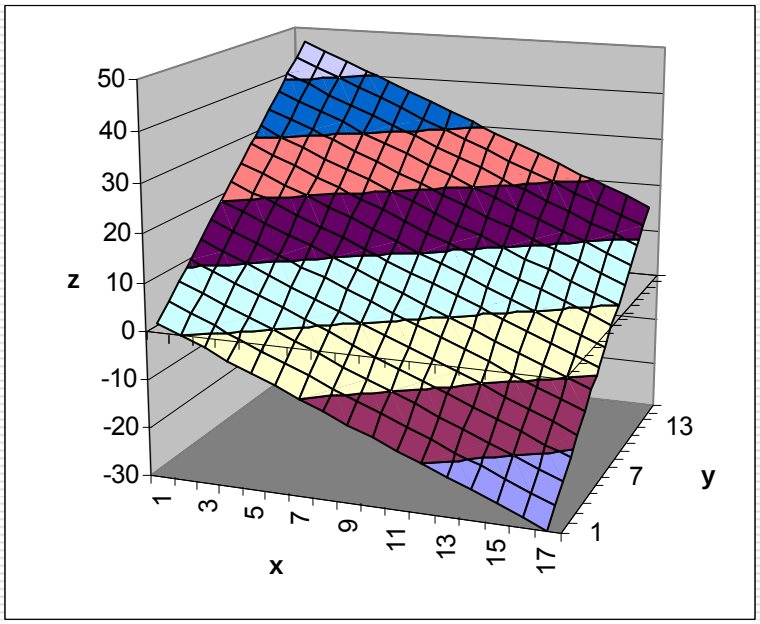
GLM

Statistična spremenljivka	Porazdelitev	Povezovalna funkcija
Zvezna	Normalna	$\mu=X\beta$
Števna	Poissonova	$\mu=\exp(X\beta)$
Dihotomna	Binomska	$\mu=\exp(X\beta)/(1+\exp(X\beta))$

Grafični prikazi v analizi povezanosti

- Razsevni diagrami
- Odgovorne površine



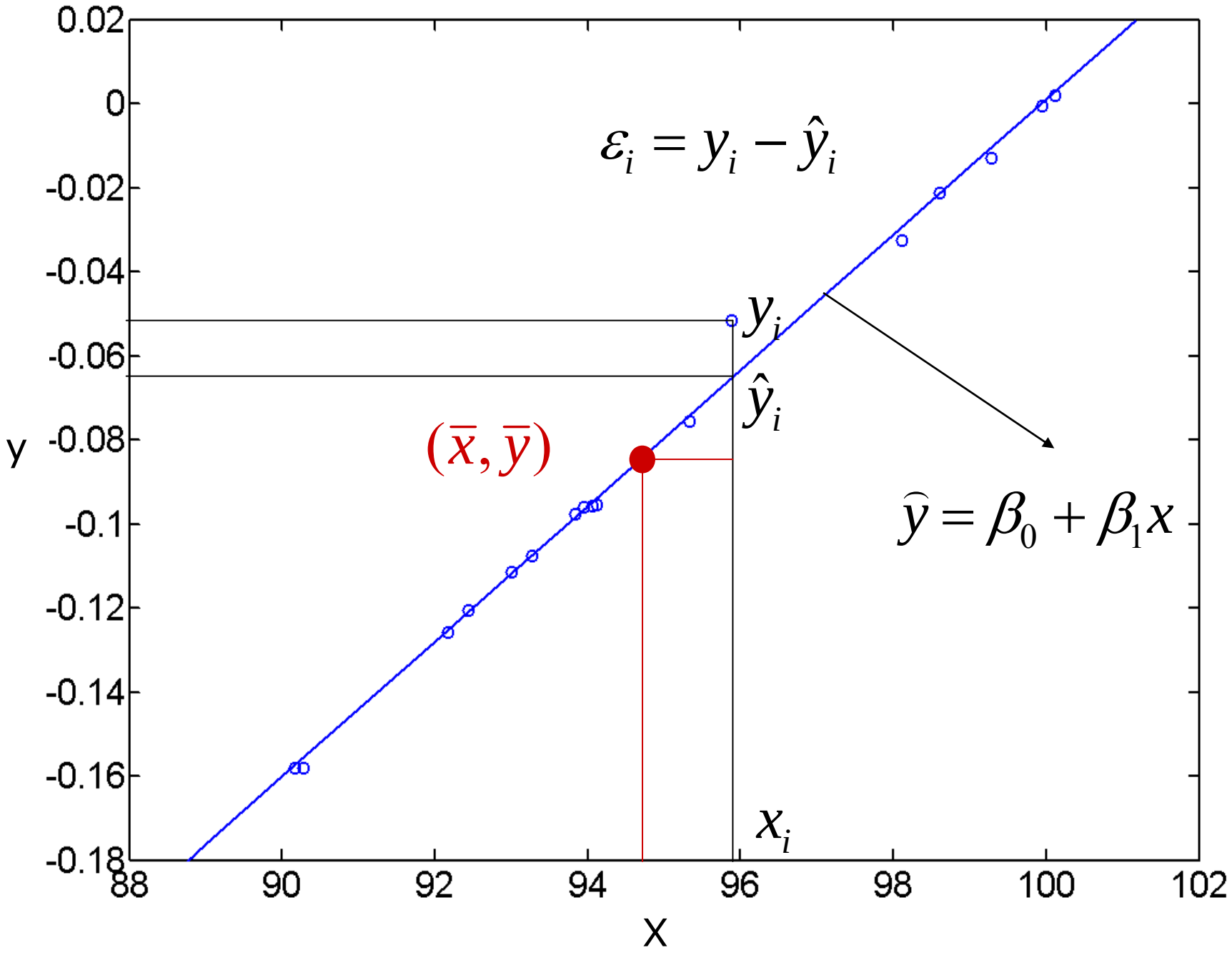


Enostavni linearni regresijski model

- Je enačba, ki opisuje odnos med odvisno (y) in neodvisno (x) spremenljivko in napako (ε).
- Enačba enostavnega linearnega regresijskega modela je:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

  - β_0 in β_1 sta parametera modela.
  - ε je napaka $N(0, \sigma_e^2)$


0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
-0.18
y
88
90
92
94
96
98
100
102
X
$\varepsilon_i = y_i - \hat{y}_i$
y_i
$\hat{y}_i$
$(\bar{x}, \bar{y})$
$\hat{y} = \beta_0 + \beta_1 x$
x_i

Metoda najmanjših kvadratov

$$\min \sum \left(y_i - \hat{y}_i \right)^2$$

y_i = opažena i-ta vrednost odvisne spremenljivke

$\hat{y}_i$ = ocena i-te vrednosti odvisne

Johann Carl Friedrich Gauss (1777-1855)

Koeficient determinacije

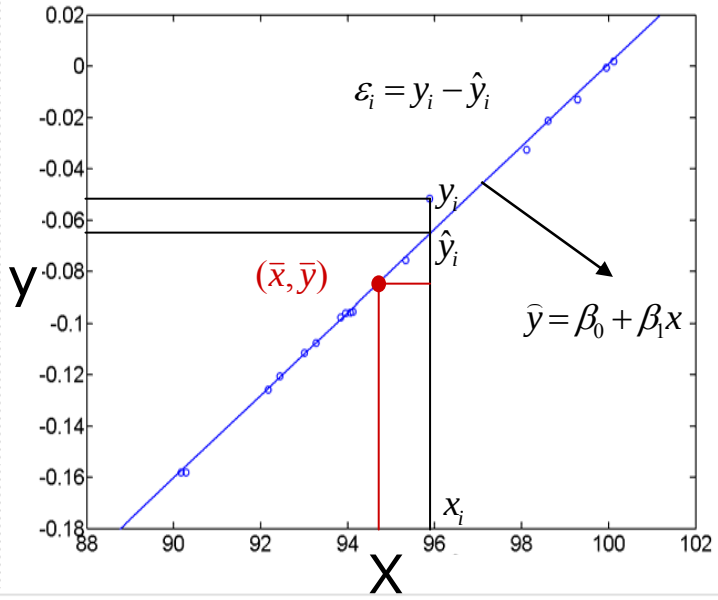
SST = SSR + SSE

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

SST = celokupna vsota kvadratov
SSR = vsota kvadratov regresije
SSE = vsota kvadratov napake

$$r^2 = \frac{SSR}{SST}$$

r – koeficient korelacije



Ocena σ_e^2

Varianca napake MSE ali s^2

s^2 = MSE = SSE/(n-2)

kjer:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$s = \sqrt{MSE}$ Standardna napaka ocene

Statistično sklepanje

- Ničelna statistična hipoteza:
- $\beta_1 = 0$
- Dva pristopa
  - *t* test
  - *F* test
- Pri obeh je potrebna ocena σ_e^2.

t test

Hipotezi

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Testna spremenljivka

$$t_{\alpha/2,n-2} = \frac{b_1}{s_{b1}} \qquad s_{b_1} = \sqrt{\frac{\text{MSE}}{\sum (x_i - \bar{x})^2}}$$

Določanje intervala zaupanja β_1

$$b_1 \pm t_{\alpha/2,n-2}\ s_{b1}$$

kjer je

b_1 točkovna ocena naklona

$t_{\alpha/2,n-2}\ s_{b1}$ kritična meja za stopnjo tveganja α

$t_{\alpha/2,n-2}$ vrednost iz t-porazdelitve z (n – 2) stopinjami prostosti

F test

Hipotezi

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Testna spremenljivka

$F_{\alpha,1,n-2} = \text{MSR/MSE}$

Previdnost pri interpretaciji statistične značilnosti

- Zavrnitev H_0: $\beta_1 = 0$ in sklep, da je povezava med spremenljivkama *x* in *y* značilna ne dopušča opredelitve povezave v smislu vzrok in posledica.

- Z zavrnitvijo H_0: $\beta_1 = 0$ nismo dokazali, da je odnos med spremenljivkama linearen.

Formule

$$L_{xx} = \sum (x_i - \bar{x})^2 = \sum \left(x_i^{\,2}\right) - \frac{\left(\sum x_i\right)^2}{n}$$

$$L_{yy} = \sum (y_i - y)^2 = \sum \left(y_i^{\,2}\right) - \frac{\left(\sum y_i\right)^2}{n}$$

$$L_{xy} = \sum \left((x_i - \bar{x})(y_i - \bar{y})\right) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$SSE = SST - SSR$$

$$SSR = \frac{L_{xy}^{\,2}}{L_{xx}}$$

$$SSE = L_{yy} - \frac{L_{xy}^{\,2}}{L_{xx}}$$

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n-k-1}$$

$$F_{k,n-k-1,\alpha} = \frac{MSR}{MSE}$$

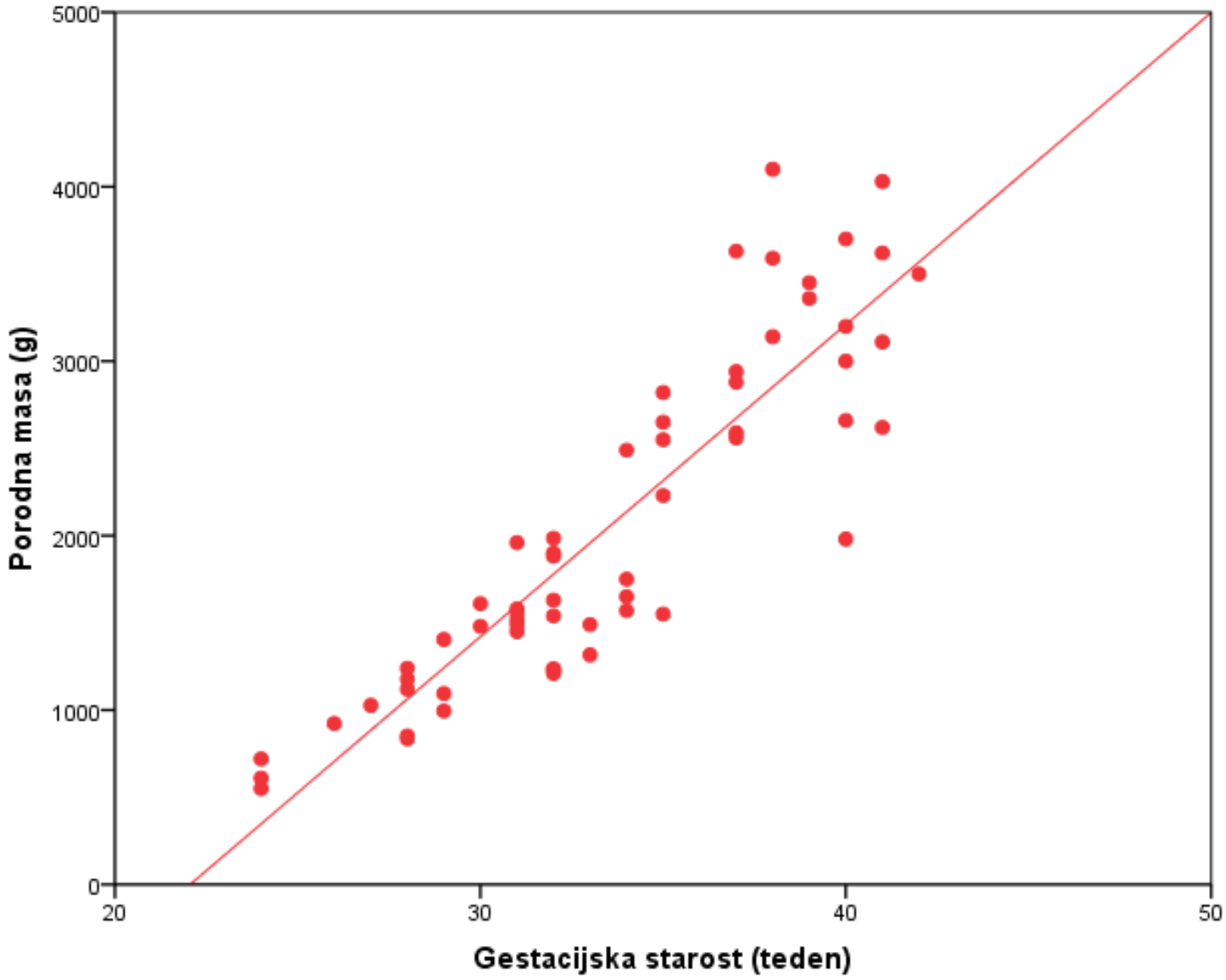
$$b_1 = \frac{L_{xy}}{L_{xx}}$$

$$se(b_1) = \sqrt{\frac{\mathrm{MSE}}{\mathrm{L_{xx}}}}$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i - b_1 \sum x_i}{n}$$

$$se(b_0) = \sqrt{\mathrm{MSE}\left(\frac{1}{\mathrm{n}} + \frac{\bar{x}^2}{L_{xx}}\right)}$$

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}}$$


5000
4000
3000
2000
1000
0
Porodna masa (g)
20
30
40
50
Gestacijska starost (teden)

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	gestacijska starost (teden)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: porodna masa (g)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.893[a]	.798	.794	437.04941

a. Predictors: (Constant), gestacijska starost (teden)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43671853	1	43671852.931	228.634	.000[a]
	Residual	11078707	58	191012.191		
	Total	54750560	59			

a. Predictors: (Constant), gestacijska starost (teden)

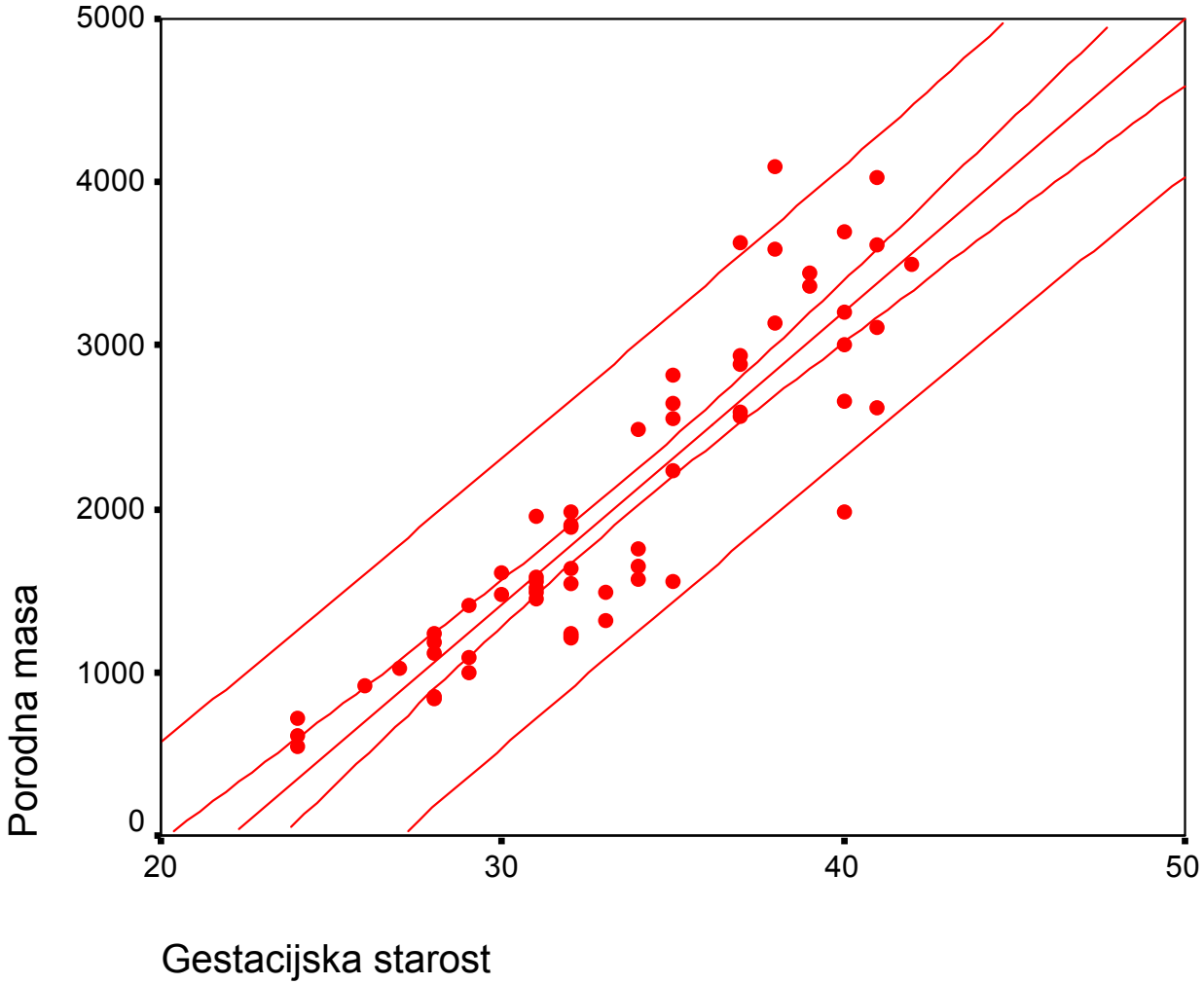
b. Dependent Variable: porodna masa (g)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-3945.614	401.102		-9.837	.000
	gestacijska starost (teden)	178.886	11.831	.893	15.121	.000

a. Dependent Variable: porodna masa (g)

Interval zaupanja in interval napovedovanja



Uporaba regresijske enačbe

- Interval zaupanja

$$\hat{y} \pm t_{n-2,\alpha/2} s \qquad s = \sqrt{MSE\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

- Interval napovedovanja

$$\hat{y} \pm t_{n-2,\alpha/2} s \qquad s = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

Multipli regresijski model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ so parametri.
- Standardizirani parametri ($x_i \rightarrow z_i$)
- ε je napaka, ki je slučajna spremenljivka.

Koeficient determinacije

- Multipli koeficient determinacije

$$R^2 = \mathrm{SSR}/\mathrm{SST}$$

- Popravljeni multipli koeficient determinacije

$$R_a^{\,2} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p-1}$$

n = 20 – 30 na posamezen prediktor (če so prediktorji med sabo korelirani je potreben večji n)

Statistično sklepanje

- Pri enostavni linearni regresiji nas *F* in *t* test vodita k istim sklepom.
- V multipli regresiji uporabimo *F* in *t* test za različna namena.

F test

- Pomembnost modela kot celote.
- Test for overall significance.

t test

- Če je izid *F* testa za model kot celota značilen, uporabimo *t* teste za ugotavljanje značilnosti vplivov posameznih neodvisnih spremenljivk.
- Za vsako neodvisno spremenljivko izvedemo en *t* test.
- Vsak t test imenujemo tudi test posamične značilnosti (test for individual significance).

F test

- Hipotezi

H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_a: Najmanj eden izmed parametrov ni enak 0

- Testna spremenljivka

F = MSR/MSE

- Zavrnitev ničelne hipoteze

Zavrni H_0 če $F > F_\alpha$

kjer je F_α odklon v F porazdelitvi s p stopinjami prostosti v števcu in (n - p – 1) stopinjami prostosti v imenovalcu ulomka.

t test

- Hipotezi

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

- Testna spremenljivka

$$t = \frac{b_i}{s_{bi}}$$

- Zavrnitev ničelne hipoteze

Zavrni H_0, če $|t| > t_{\alpha/2}$

kjer je $t_{\alpha/2}$ odklon v t porazdelitvi z $n - p - 1$ stopinjami prostosti.

Atributivne spremenljivke

- Rekodiranje s slepimi spremenljivkami
- k nivojev osnovne atributivne spremenljivke nadomestimo s (k-1) slepimi spremenljivkami

Primer:
Vpliv kajenja na pljučno funkcijo

Z raziskavo želimo opredeliti vpliv kajenja na pljučno funkcijo. Naključno smo izbrali 1000 ljudi obeh spolov in spremljali pljučno funkcijo (parameter FEV), poleg tega smo beležili še njihovo starost in telesno višino. Za analizo rezultatov raziskave smo uporabili metodo multiple regresije. Neodvisne spremenljivke v analizi so bile:

starost [leta]
telesna višina [cm]
spol (0 = ženski, 1 = moški)
kajenje (0 = nekadilec, 1 = kadilec)
Odvisna spremenljivka pa je bila FEV [liter].

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.497[a]	.247	.243	.25269

a. Predictors: (Constant), kajenje, t. višina (cm), starost (leta), spol

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20.786	4	5.196	81.383	.000[a]
	Residual	63.533	995	.064		
	Total	84.319	999			

a. Predictors: (Constant), kajenje, t. višina (cm), starost (leta), spol

b. Dependent Variable: FEV (l)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.666	.118		22.568	.000
	starost (leta)	-1.9E-03	.001	-.063	-2.296	.022
	t. višina (cm)	-5.6E-04	.001	-.024	-.844	.399
	spol	.219	.017	.376	13.208	.000
	kajenje	-.275	.023	-.327	-11.857	.000

a. Dependent Variable: FEV (l)

$$y = b_0 + b_1\ starost + b_2\ višina + b_3\ spol + b_4\ kajenje$$

Primer: Pravastatin

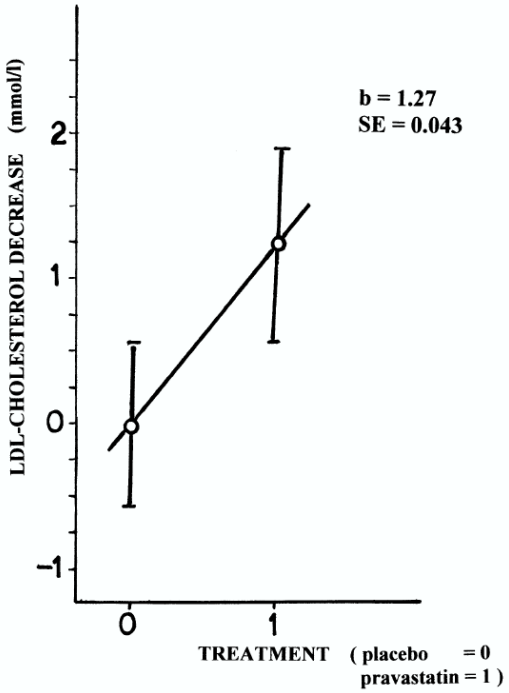
- Randomizirana klinična raziskava
- Paralelni načrt (placebo, n=434 in pravastatin, n=438)
- Meritev LDL pred zdravljenjem in 2 leti po začetku zdravljenja
- Primarni izid je znižanje vrednosti LDL

Jukema JW, Bruschke AV, van Boven AJ, Reiber JH, Bal ET, Zwinderman AH, et al. Effects of lipid lowering by pravastatin on progression and regression of coronary artery disease in symptomatic men with normal to moderately elevated serum cholesterol levels. The Regression Growth Evaluation Statin Study (REGRESS). Circulation 1995;91:2528-40.

Rezultati

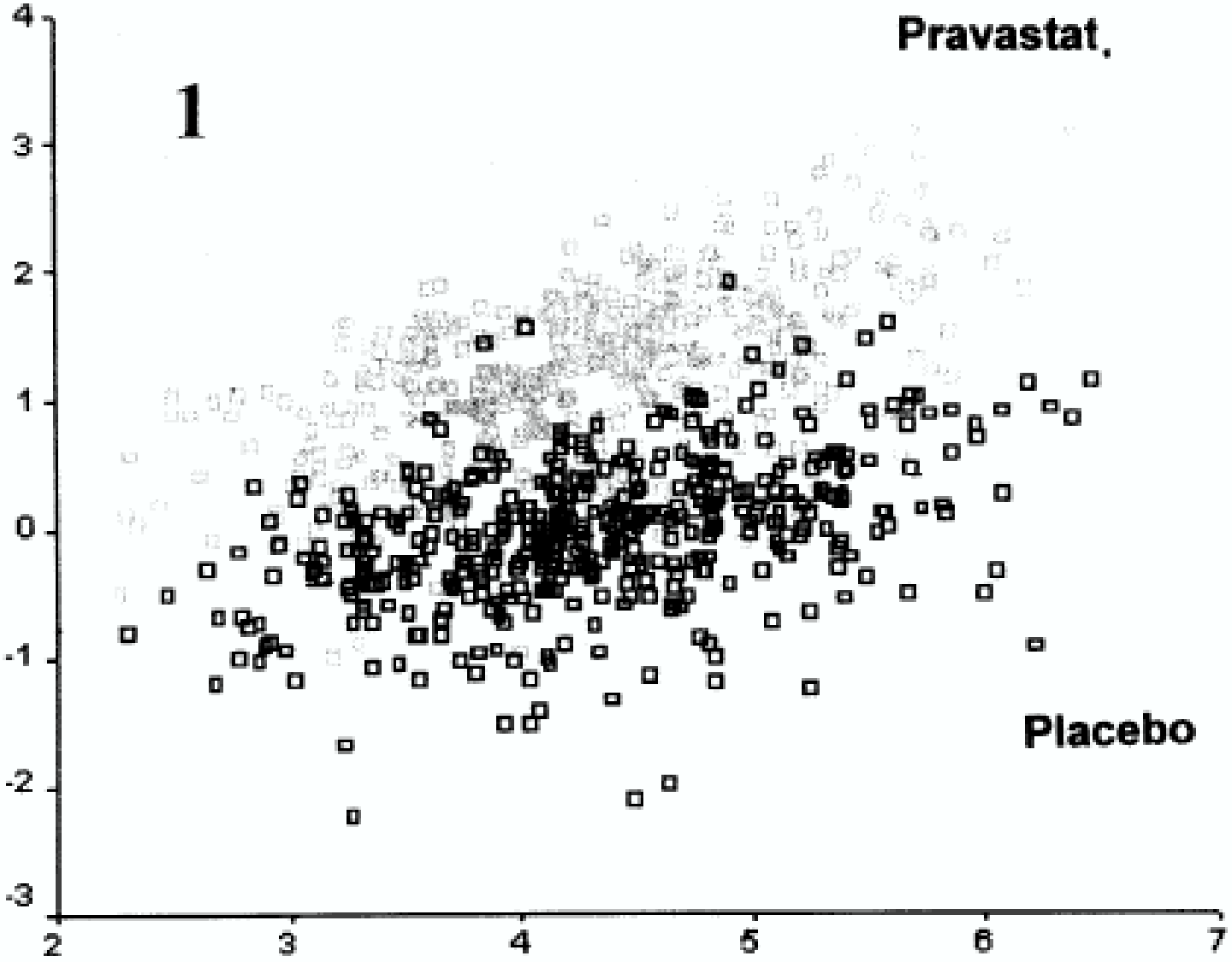
Placebo: LDL=-0.04±0.59 mmol/L

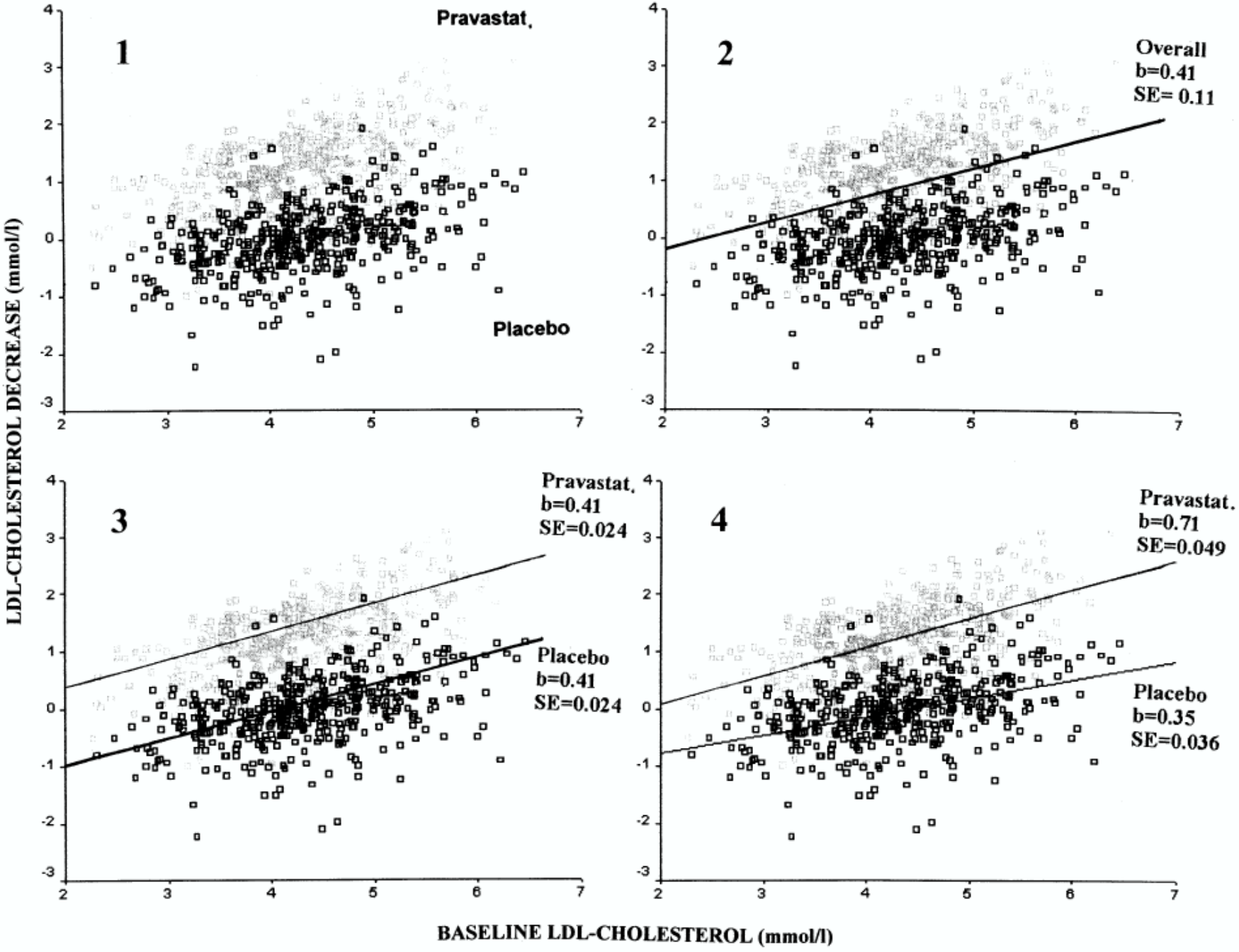
Pravastatin: LDL=1.23±0.68 mmol/L



Δ=1.23-(-0.04)=1.27 mmol/L

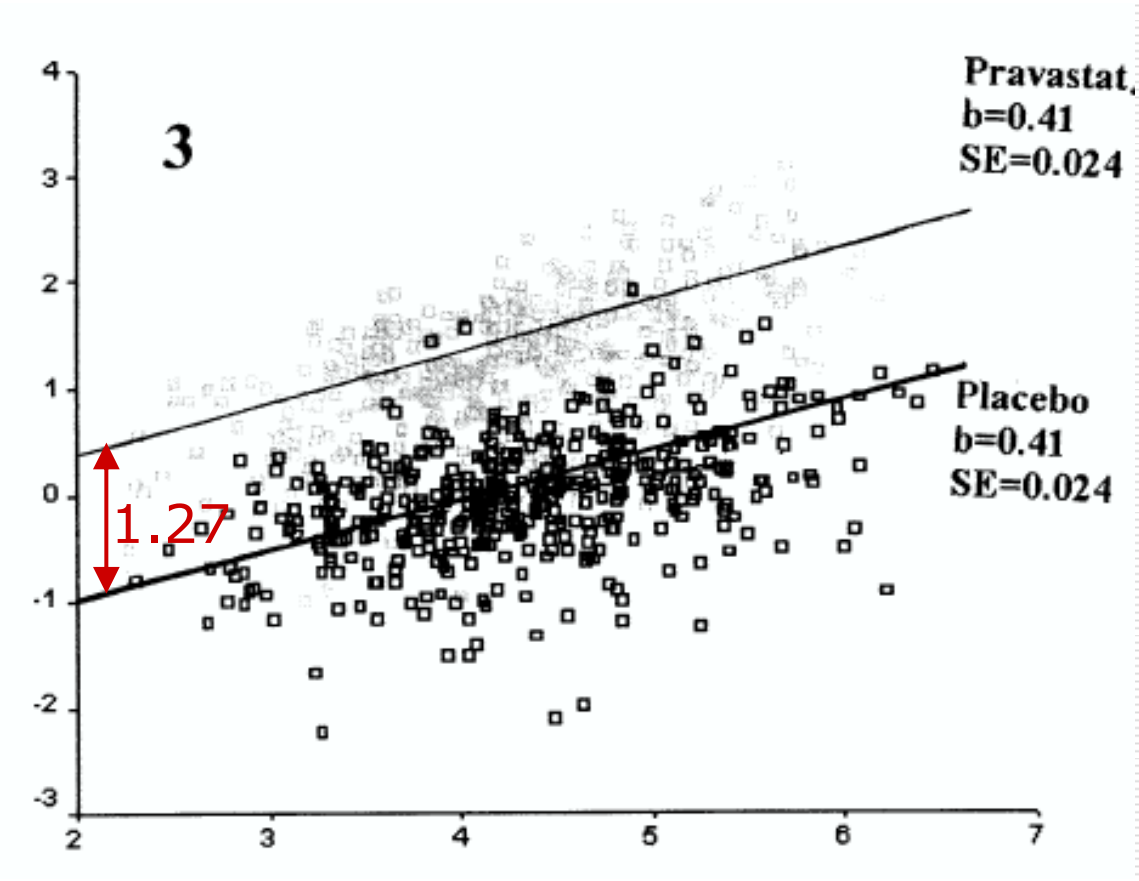
$$se = \sqrt{\left(0.68^2 / 438\right) + \left(0.59^2 / 434\right)} = 0.043 \quad mmol / L$$


1
Pravastat.
Placebo
4
3
2
1
0
-1
-2
-3
2
3
4
5
6
7
LDL-cholesterol decrease (mmol/l)
Baseline LDL-cholesterol (mmol/l)


1
Pravastat.
Placebo
2
Overall
b=0.41
SE= 0.11
3
Pravastat.
b=0.41
SE=0.024
Placebo
b=0.41
SE=0.024
4
Pravastat.
b=0.71
SE=0.049
Placebo
b=0.35
SE=0.036
LDL-CHOLESTEROL DECREASE (mmol/l)
BASELINE LDL-CHOLESTEROL (mmol/l)

$y=b_0+b_1x_1+b_2x_2$

$x_1=1$ (pravastatin), $x_1=0$ (placebo)
x_2=baseline



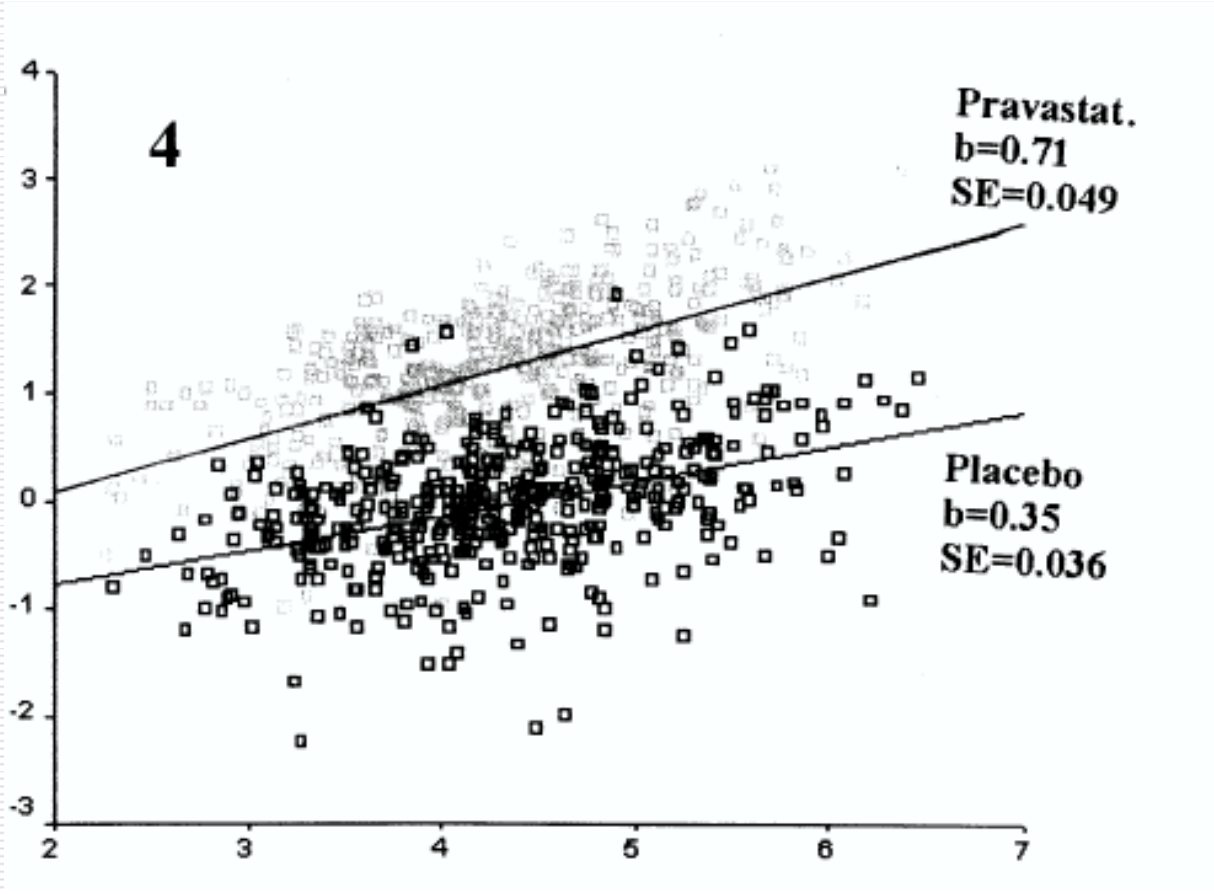
Placebo
$y=b_0+b_2x_2$

Pravastatin
$y=b_0+b_1+b_2x_2$

$b_1=1.27\pm0.037$

Izboljšana natančnost se: 0.043 → 0.037

$\Delta b=0.36$ (se=0.06; $p<0.0001$)



Tveganje, obeti ter razmerja

Relativno tveganje

ang. Relative Risk

$$RR = \frac{p_1}{p_2}$$

Razmerje obetov

ang. Odds Ratio

$$\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

Povezava med razmerjem obetov ter relativnim tveganjem:

$$\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1}{p_2} \times \frac{1-p_2}{1-p_1} = RR \times \frac{1-p_2}{1-p_1}$$

Preprosta logistična regresija

- Opazovani pojav= binarna spremenljivka 0/1
- Kako uporabiti linearni model za opisovanje odnosa?

Preprosta logistična regresija

Vrednost binarne spremenljivke zapisati kot:

- Verjetnost, da zavzame vrednost 1 pri danem X:
  - π (x)- populacija
  - p(x)- vzorec
  - Zavzame le vrednosti med 0 in 1- preozek razpon
- Razmerje verjetnosti, da dogodek zgodi in da se ne zgodi:

$$\frac{p(x)}{1-p(x)}$$

  - Zavzame vrednosti med 0 in + neskončnostjo

Preprosta logistična regresija

- Logaritmiranje=> med – in + neskončnostjo

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = \log it$$

Populacija:

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \text{logit}[p(x)] = \alpha + \beta x$$

Ocena parametrov iz vzorca:

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = a + bx$$

$$p(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$$

Verjetnost, da dogodek zgodil pri danem X

Pomen regresijskega koef. b

- Lin. regresija: $b = y_{(x+1)} - y_{(x)}$

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = a + bx = \log it$$

- Log. regresija: $b = \log it_{(x+1)} - \log it_{(x)}$
- Binarni X (0/1) => x=0, x+1=1.
- Obet da bo y=1:

$$O_{x=1} = \frac{p(1)}{1-p(1)} \qquad O_{x=0} = \frac{p(0)}{1-p(0)}$$

- Logaritmiranje:

$$\log it_{x=1} = \ln\left[\frac{p(1)}{1-p(1)}\right] \qquad \log it_{x=0} = \ln\left[\frac{p(0)}{1-p(0)}\right]$$

Pomen regresijskega koef. b

- Log. razmerja obetov:

$$RO = \frac{O_{x=1}}{O_{x=0}} = \frac{\left[\frac{p(1)}{1-p(1)}\right]}{\left[\frac{p(0)}{1-p(0)}\right]} \qquad O_{x=1} = \frac{p(1)}{1-p(1)} \qquad O_{x=0} = \frac{p(0)}{1-p(0)}$$

$$\ln RO = \ln \frac{\left[\frac{p(1)}{1-p(1)}\right]}{\left[\frac{p(0)}{1-p(0)}\right]} = \log it(1) - \log it(0)$$

$$\ln RO = b = \log it_{(1)} - \log it_{(0)}$$

$$RO = e^{b}$$

Ocenjevanje b v vzorcu

- p(x)? =>b in a ?

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = a + bx = \log it$$

- Lin. regresija: metoda najmanjših kvadratov ostankov.
- Log. regresija: metoda največjega verjetja (maximum likelihood method).
  - Funkcija največjega verjetja
  - Oz. logaritem funkcije verjetja ("log likelihood"):
    - Nelinearna funkcija parametrov modela a in b.
    - Iteracijska metoda, več ocen parametrov. Nove ocene, dokler še zveča funkcijo največjega verjetja. Lokalna/globalna točka največjega verjetja. Start?

Ocenjevanje b v populaciji

- Vzorčna porazdelitev b- ja: normalna porazdelitev = > interval zaupanja:

$$\beta = b \pm z * SE_{(b)}$$

- Vzorčna ocena razmerja obetov: nenormalna, nesimetrična porazdelitev
- Interval zaupanja za RO iz b (spodnjo/zgornjo mejo):

$$e^{b \pm z * SE_b}$$

Multipla logistična regresija

- Matematični model:

$$p(x) = \frac{e^{a+b_1x_1+b_2x_2+\ldots+b_kx_k}}{1+e^{a+b_1x_1+b_2x_2+\ldots+b_kx_k}}$$

- b(k): sprememba v logitu, ki spremlja spremembo X(k) za 1 enoto, medtem ko se ostale sprem. ne spreminjajo.
- Regresijski koef. b = logaritem razmerja obetov
- Razmerje obetov= antilogaritem b: e^b
- Metoda največjega verjetja

Testiranje pomembnosti modela kot celote

- Ali se log verjetja modela s spremenljivkami X statistično značilno poveča v primerjavi z log verjetja modela brez njih.
- G- statistika- test razmerja dveh verjetij:

$$G = 2\ln\left[\frac{verjetje_{SPREM.X}}{verjetje_0}\right]$$

- Porazdeljuje po hi- kvadrat, df: k
- k: št. vseh spremenljivk v modelu
- V SPSS-u: “Omnibus test”

Vrednotenje prileganja modela

- Nagelkerkejev R^2

- Test hi- kvadrat
- Hosmer- Lemeshov test:
  - Oba primerjata opazovano število enot, pri katerem se je opazovani dogodek zgodil, s pričakovanim številom, ki temelji na enačbi logistične regresije.

Testiranje pomembnosti b

- Ho: b=0, Ha: b≠ 0
- Testi:
  - Z: $$z = \frac{b - 0}{SE_b}$$
  - Waldova statistika: $$\chi^2 = \left(\frac{b}{SE_b} \right)^2$$
  - G- statistika: $$G = 2 \ln \left[\frac{verjetje_x}{verjetje_0} \right]$$

G- statistika: Model, ki vključuje X, pove več o Y.