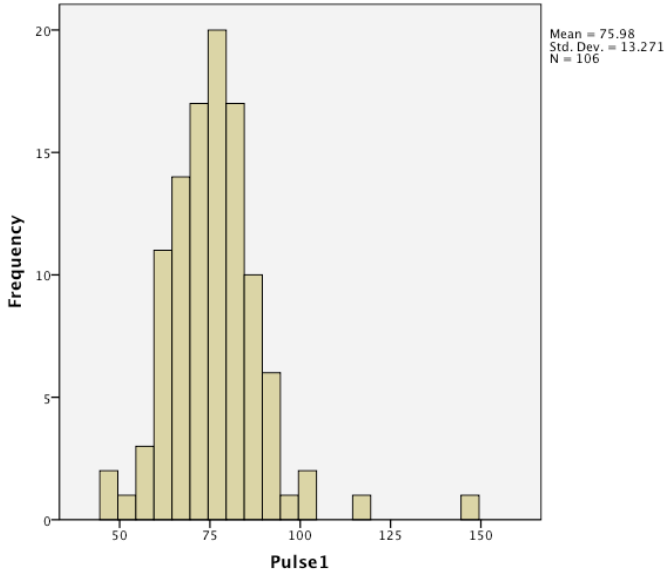
Korelacija in linearna regresija

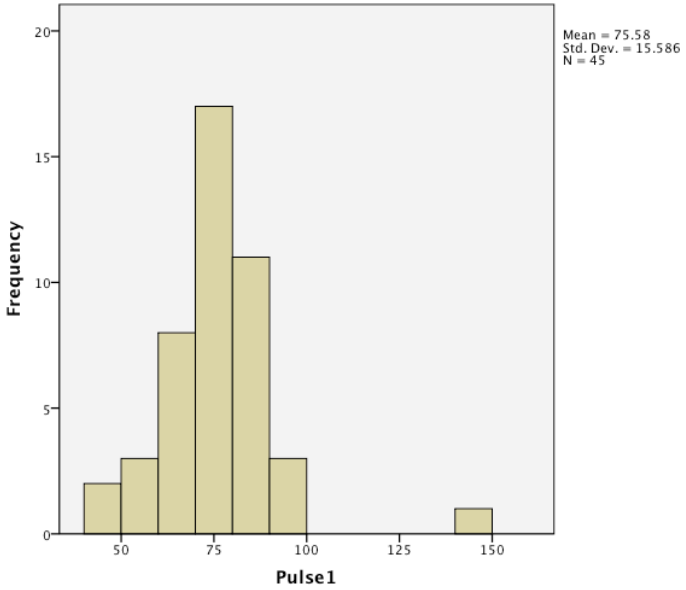
Nataša Kejžar

Inštitut za biostatistiko in medicinsko informatiko
Medicinska fakulteta, Univerza v Ljubljani

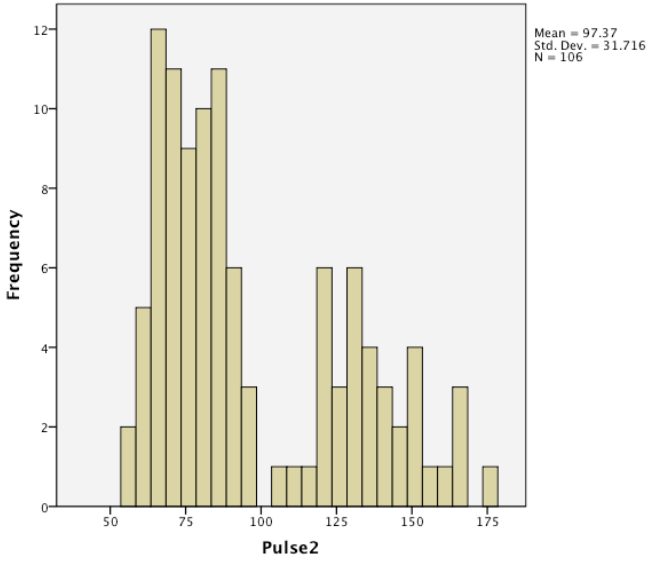
Natasa.Kejzar at mf.uni-lj.si

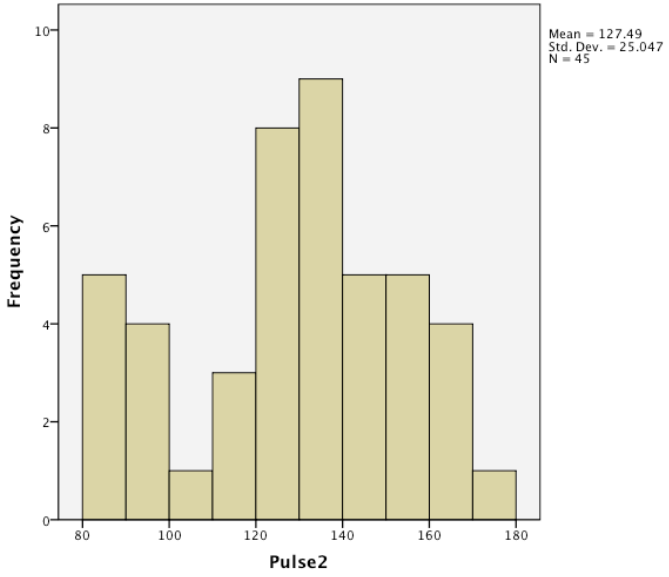
Pulz pred obremenitvijo





Pulz po obremenitvi





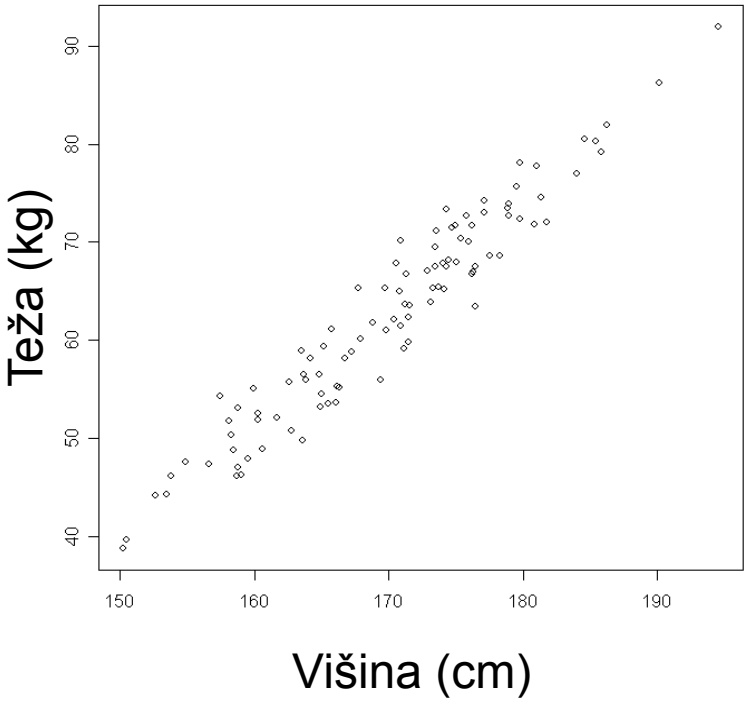
Korelacija

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

- Ali obstaja linearna povezanost med dvema številskima spremenljivkama?

Razsevni diagram (scatterplot)

Vzorčni korelacijski koeficient: r=0,96

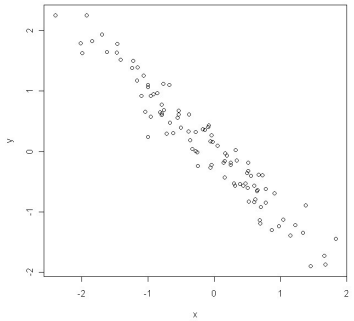


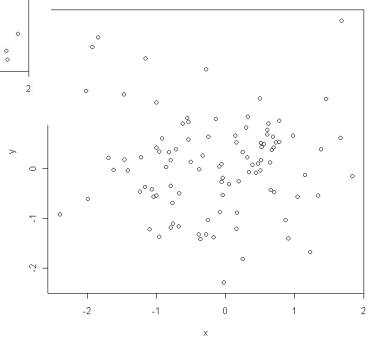
-1≤r≤1

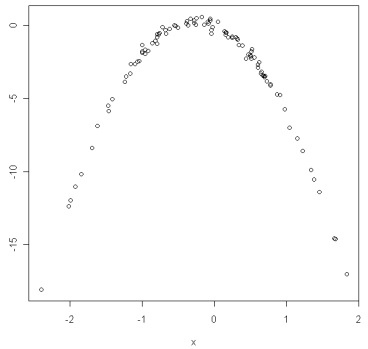
Negativna linearna povezanost

Pozitivna linearna povezanost

r=0: ni linearne povezanosti

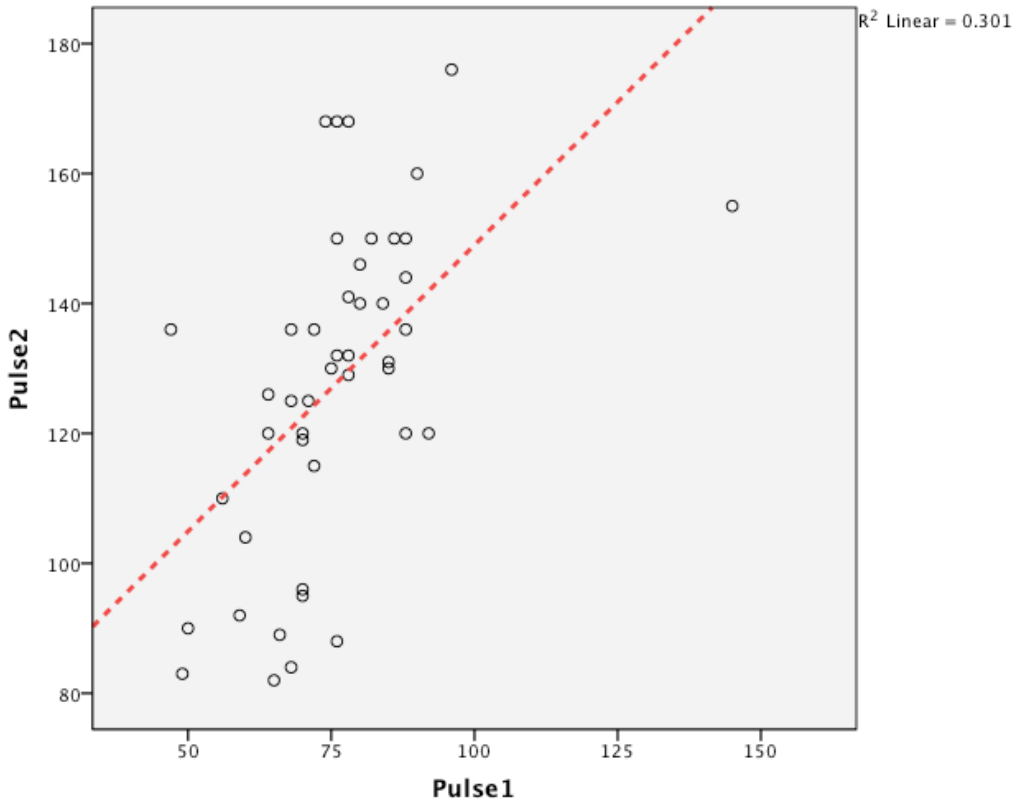






Ali sta lahko pulza linearno povezana?

Ali sta (linearno) povezana?



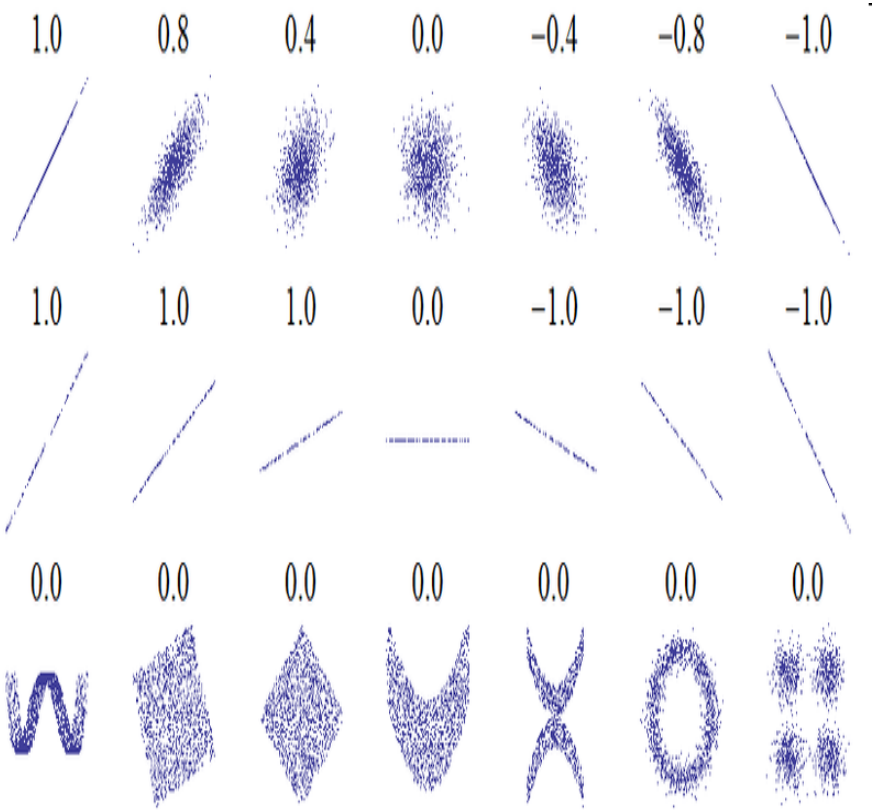
r = 0,548

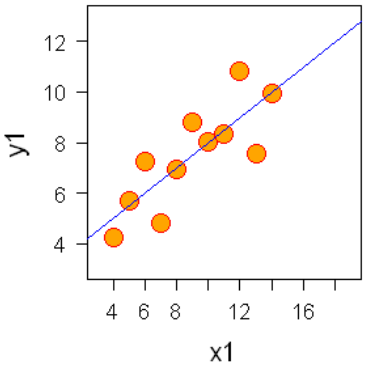
95% IZ od 0,379 do 0,756

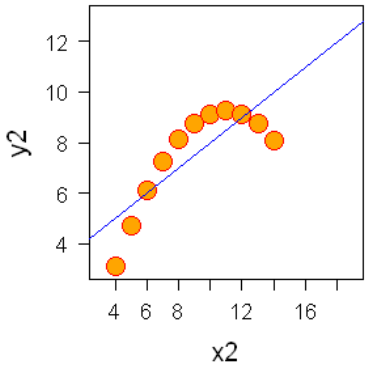
Nekaj primerov

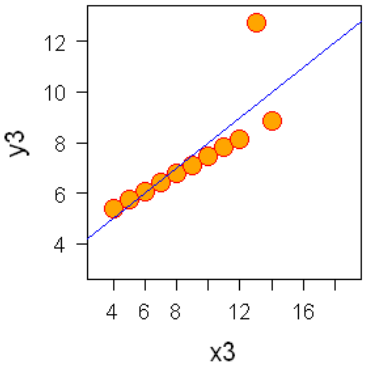
"The Anscombe quartet"

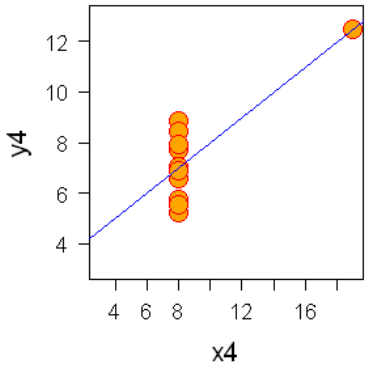
r=?











www.wikipedia.org, "Correlation and dependance"

Moramo si "ogledati/narisati" podatke!

Regresijski modeli

- Statistični modeli, ki poskušajo pojasniti nek izid z uporabo drugih spremenljivk

Izid	**Vrsta**	**"Druge spremenljivke"**	**Model**
Teža	Številska	Višina	Linearna regresija
Teža	Številska	Višina, spol, izobrazba	Multipla linearna regresija
Kajenje (Da/Ne)	Dihotomka	Starost, spol, izobrazba	Logistična regresija
Čas do smrti (dogodka)	Številska + dihotomka	Zdravilo, spol, mutacije	Analiza preživetja

Model, ki ga bomo uporabljali je odvisen od vrste izida

S pomočjo regresije lahko: testiramo povezanost med izidom in drugimi spremenljivkami ter predvidevamo izid, če je vrednost drugih spremenljivk znana

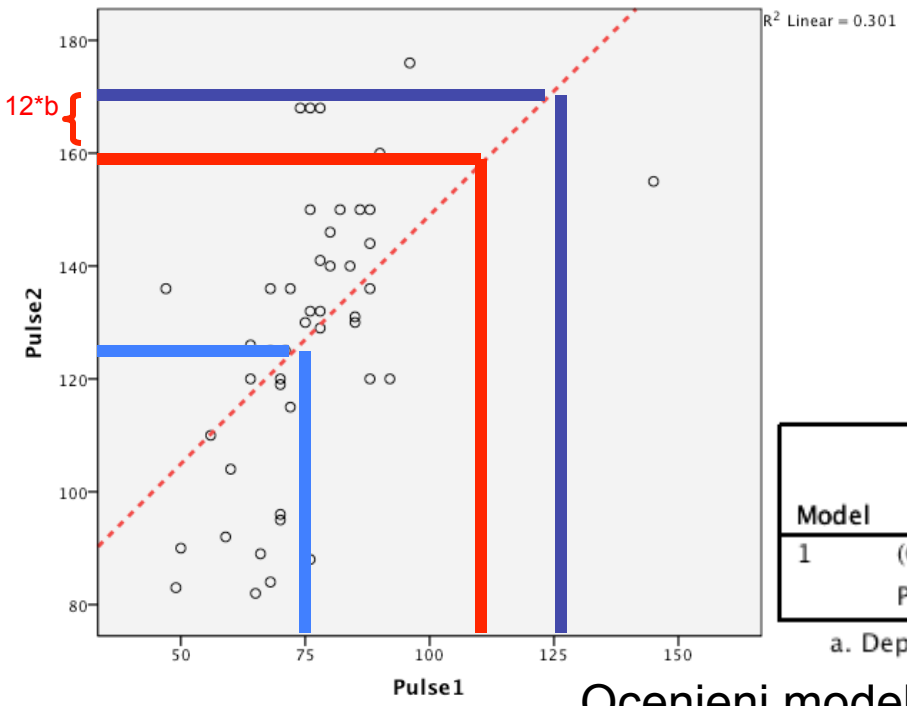
Terminologija

Outcome	"Other variables"
Izid (Response variable)	Pojasnjevalne spremenljivke
Outcome variable	Covariables, covariates
Odvisne	Neodvisne
Y	X

Ali lahko napovemo pulz po obremenitvi na podlagi pulza pred obremenitvijo

Ali sta (linearno) povezana?

pulzPo= α + β*pulzPred + ε



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.548[a]	.301	.284	21.188

a. Predictors: (Constant), Pulse1
b. Dependent Variable: Pulse2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8298.657	1	8298.657	18.485	.000[a]
	Residual	19304.587	43	448.944		
	Total	27603.244	44			

a. Predictors: (Constant), Pulse1
b. Dependent Variable: Pulse2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	60.896	15.808		3.852	.000
	Pulse1	.881	.205	.548	4.299	.000

a. Dependent Variable: Pulse2

pulzPo=a+ b*pulzPred

a : ocena za α 60,9

b: ocena za β 0,88

Ocenjeni model: pulzPo= 60,9 + 0,88*pulzPred

Ocenjeni pulzPo, ki je imel pulzPred 75?

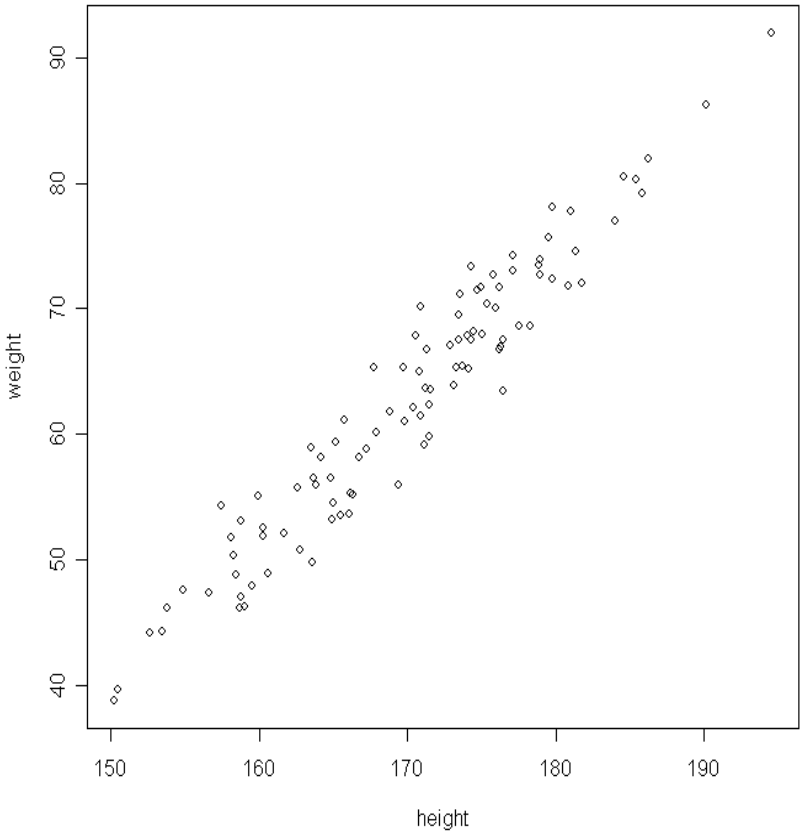
Ocenjeni pulzPo, ki je imel pulzPred 125?

Razlika za pulzPo za dva študenta, ki imata za 12 udarcev pulza pred obremenitvijo razlike?

Linearna regresija

Ali lahko napovemo vrednost teže na podlagi višine?

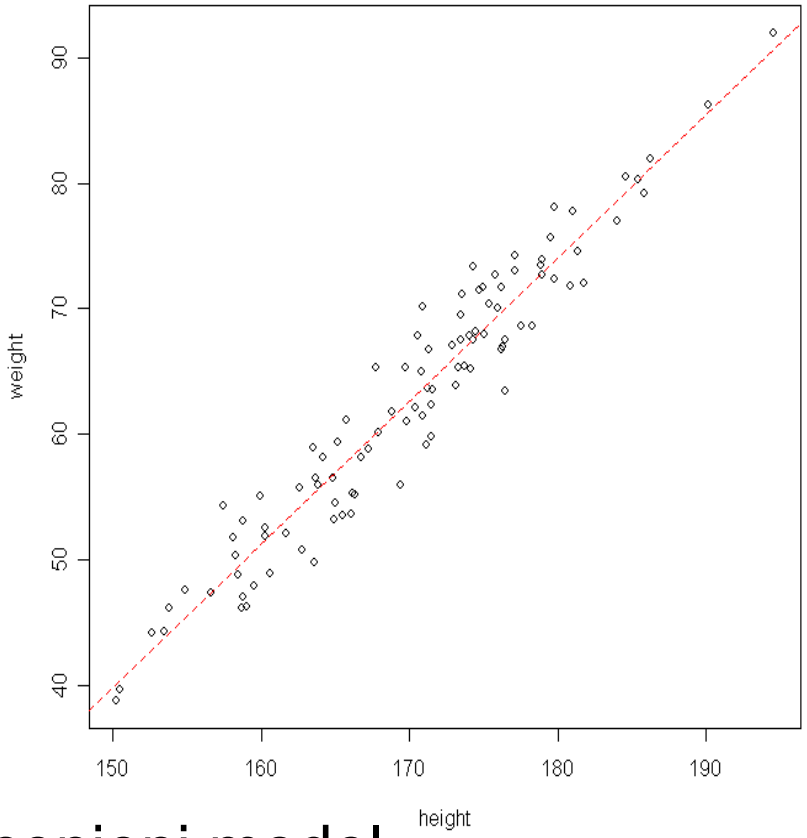
Teža=α+βVišina



Odvisna spremenljivka (Y)

Neodvisna spremenljivka (X)

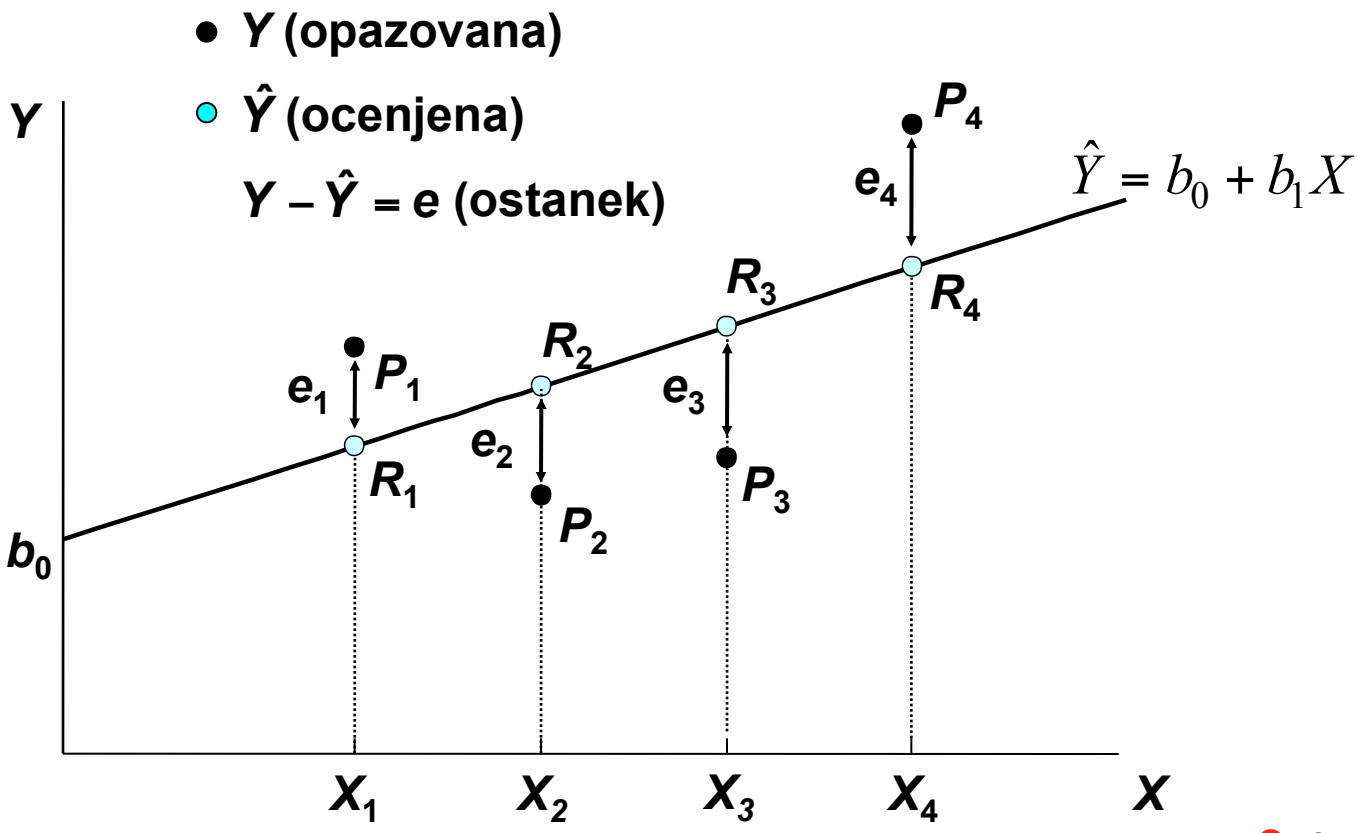
Iščemo “najboljšo premico”



Ocenjeni model

Teža=-131,54+1,142*Višina

Kako oceniti najboljšo premico



Metoda najmanjših kvadratov: minimizira Σe^2 (torej varianco ostankov $\Sigma e^2/(n-2)$) Povprečje ostankov =0!

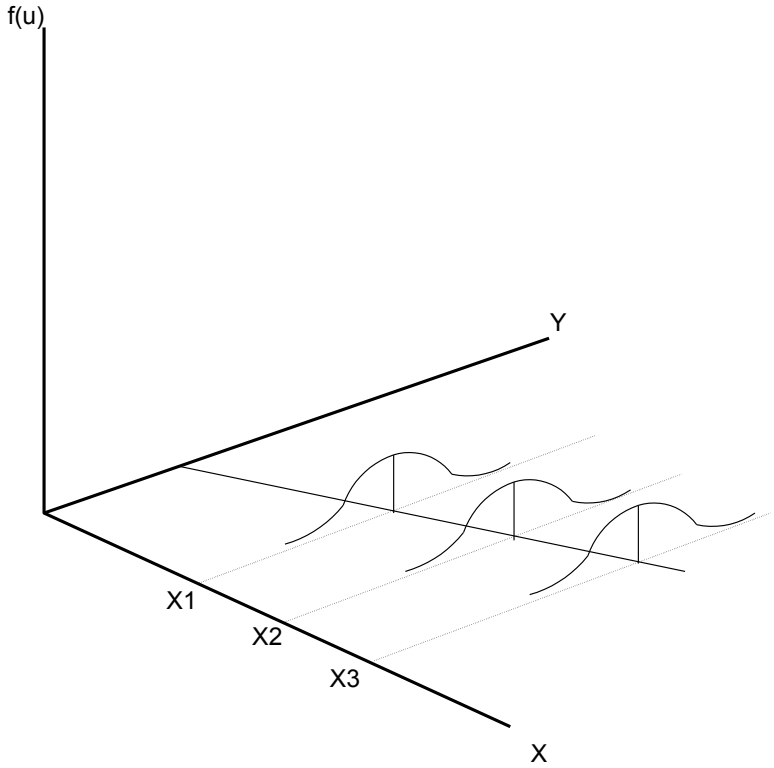
Model

$$y=\alpha+\beta x+\varepsilon$$

Teža=α+βVišina+ε

PREDPOSTAVKE

1. Razmerje med Y in X je linearno
2. Vrednosti Y so normalno porazdeljene okrog regresijske premice (za vsako vrednost od X)
3. Razpršenost Y je konstantna okrog regresijske premice (za vsako vrednost X)
4. Opazovanja so neodvisna



Predpostavke moramo preveriti (za nekatere moramo najprej oceniti model!)

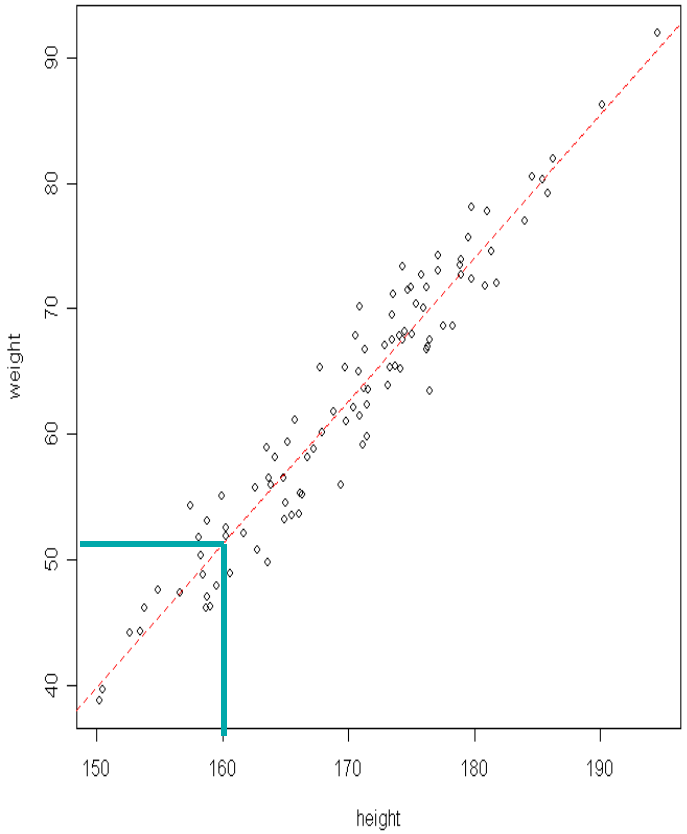
Interpretacija ocenjenega modela

Teža=α+βVišina+ε : **Populacija**

Teža=a+bVišina : Ocena z vzorca

Teža=-131,54+1,14*Višina : Ocena

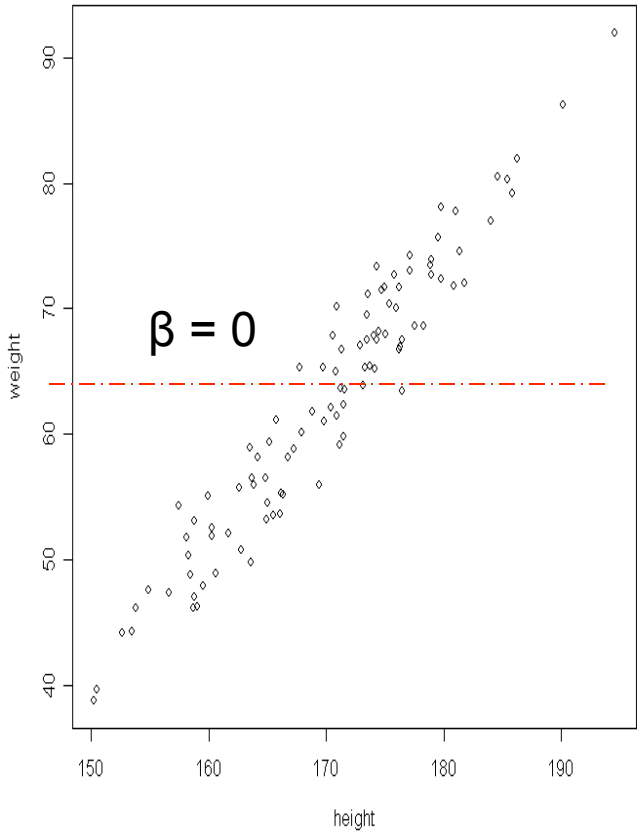
- a (regresijska konstanta)
  - Geometrično: presečišče z osjo y
  - Ocenjena vrednost odvisne spremenljivke, ko je neodvisna spremenljivka 0.
  - Teža, ko je višina=0 (v tem primeru nima smisla)
- b (regresijski koeficient)
  - Geometrično: naklon premice
  - Ocenjena razlika odvisne spremenljivke, ko se neodvisna spremenljivka razlikuje za 1 enoto
  - povrečna razlika v teži za dve osebi, ki imata 1 enoto razlike v višini (v tem primeru: 1.14 kg za 1 cm razlike)
- Kolikšno težo pričakujmo za osebo, ki ima 160 cm?
  - -131,54+1,14*160=50,82



- Najpomembnejše vprašanje : ali lahko sklepamo, da je β≠0?
  - H0: β=0

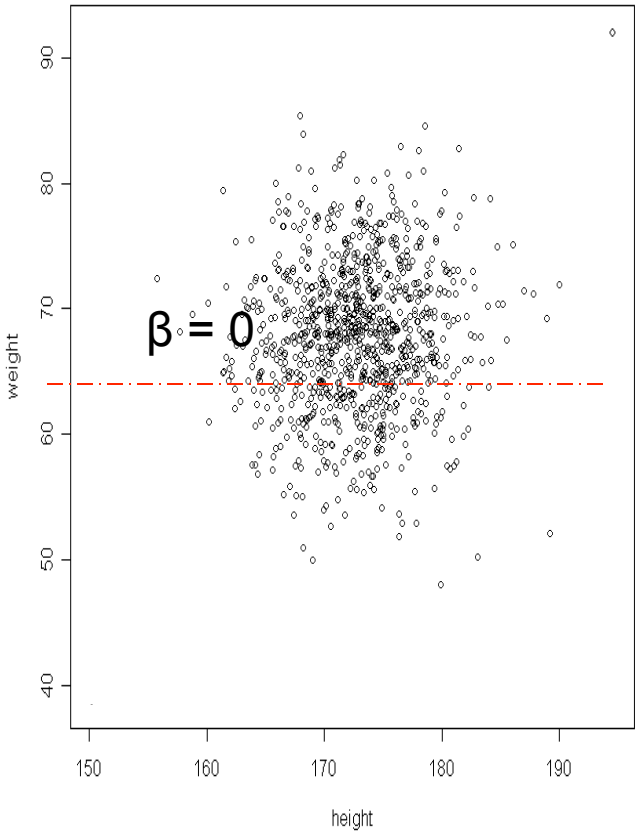
Interpretacija ocenjenega modela

- Najpomembnejše vprašanje : ali lahko sklepamo, da je β≠0?
  - H0: β=0



Interpretacija ocenjenega modela

- Najpomembnejše vprašanje : ali lahko sklepamo, da je β≠0?
  - H0: β=0



Kaj dobimo, ko program oceni model?

Dobili smo (z linearno regresijo): Teža=-131.54+1.14*Višina

Residuals: - **OSTANKI**

Min	1Q	Median	3Q	Max
-0.7852	-0.5134	-0.1736	0.3360	1.4119

Coefficients:

Ocena

Standardna napaka

Testna statistika, t=b/SE(b)

$t \sim t_{n-2}$

P-vrednost

		Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	a	-131.54354	5.36023	4.54	<2e-16 ***
višina	b	1.14217	0.03148	36.28	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.839 on 98 degrees of freedom

Multiple R-Squared: 0.93071, Adjusted R-squared: 0.93

F-statistic: 1317 on 1 and 98 DF, p-value: < 2.2e-16

Mera pojasnjene variabilnosti

$$b = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$t = \frac{b - \beta}{SE(b)} \sim t_{n-2}$$

$H0: \beta = 0$ običajna H0

$$SE(b) = \sqrt{\frac{\sum_{i=1}^{n} e^2{}_i / (n-2)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

PONOVITEV:
95% interval zaupanja za razliko povprečij

95% interval zaupanja za $\mu_1 - \mu_2$

Interval za katerega imamo 95% zaupanje, da vsebuje neznano populacijsko razliko povprečij

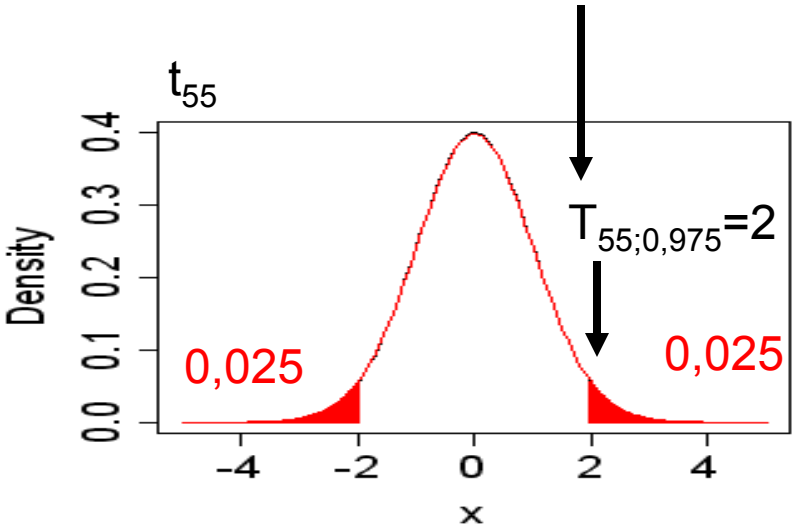
$$95\%\ \text{IZ za}\ \ \mu_{G1} - \mu_{G2}:$$

$$(\bar{x}_{G1} - \bar{x}_{G2}) - t_{55;1-0,05/2} \cdot SE, (\bar{x}_{G1} - \bar{x}_{G2}) + t_{55;1-0,05/2} \cdot SE$$

$$\bar{x}_{G1} - \bar{x}_{G2} = 9 \quad SE = 4,19$$

$$9 - 2 \cdot 4,19 \text{ do } 9 + 2 \cdot 4,19$$

$$0,62 \text{ do } 17,38$$



S 95% zaupanjem lahko trdimo, da je razlika populacijskih povprečij

($\mu_{G1} - \mu_{G2}$) med 0,62 in 17,38

$$t_{df,1-\alpha/2} : P(t_{df} \le 1 - \frac{\alpha}{2}) = 1 - \alpha$$

(1-α)% interval zaupanja za regresijske koeficiente

$$\{b - t_{df,\,1-\alpha/2} * SE(b) \;,\; b + t_{df,\,1-\alpha/2} * SE(b)\}$$

$$\{a - t_{df,1-\alpha/2} * SE(a) \;,\; a + t_{df,\,1-\alpha/2} * SE(a)\}$$

df: n-2

velikost vzorca – število ocenjenih koeficientov

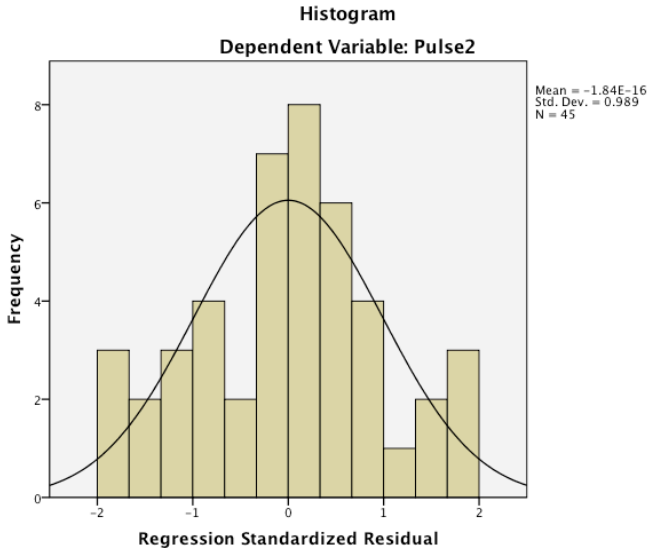
α presečišče **α** : 95% CI (-142.18, -120.91)

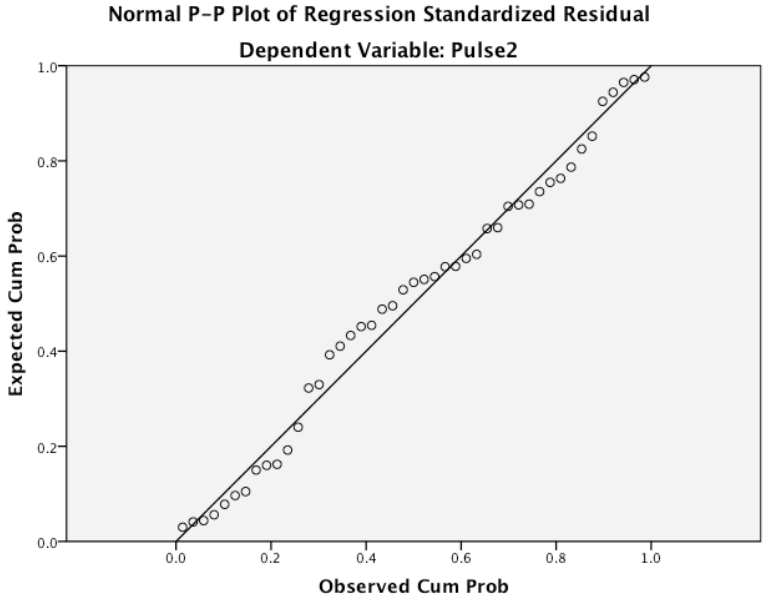
β višina : 95% CI (1.08, 1.205)

Preverjanje predpostavk

$$y=\alpha+\beta x+\varepsilon$$

1. Razmerje med Y in X je linearno
2. Vrednosti Y so normalno porazdeljene okrog regresijske premice (za vsako vrednost od X)
3. Razpršenost Y je konstantna okrog regresijske premice (za vsako vrednost X)
4. Opazovanja so neodvisna

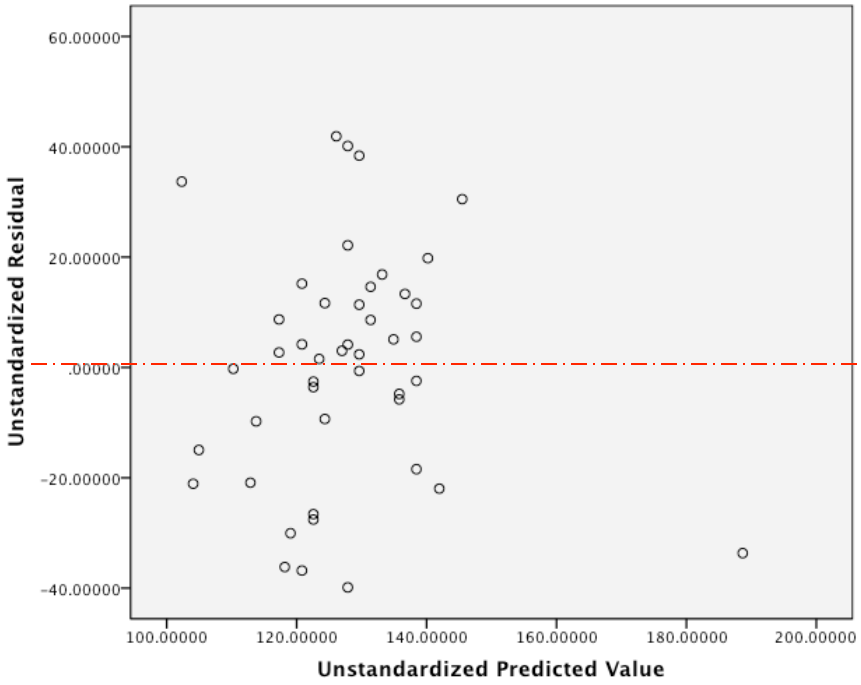




Preverjanje predpostavk

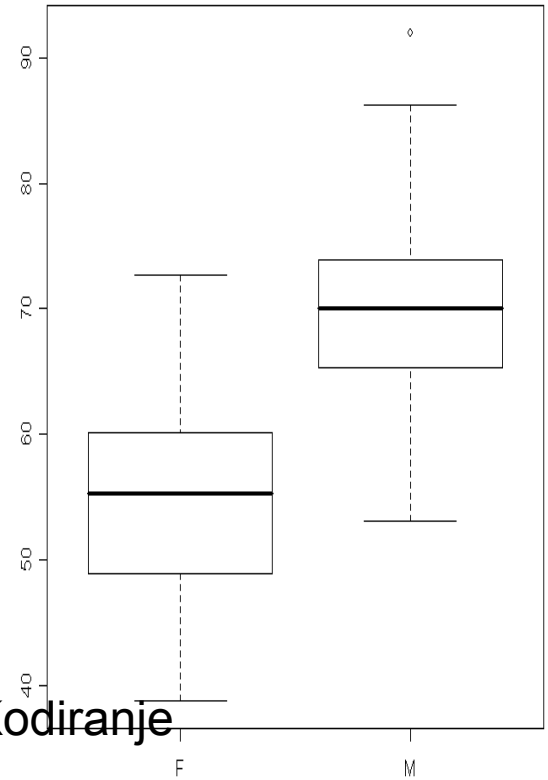
$$y=\alpha+\beta x+\varepsilon$$

1. Razmerje med Y in X je linearno
2. Vrednosti Y so normalno porazdeljene okrog regresijske premice (za vsako vrednost od X)
3. Razpršenost Y je konstantna okrog regresijske premice (za vsako vrednost X)
4. Opazovanja so neodvisna



Uporaba opisnih spremenljivk v linearni regresiji

Weight= $\alpha + \beta$ *gender+ ε



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**55.269**	**1.098**	**50.318**	**< 2e-16**	***
genderM	**14.814**	**1.553**	**9.537**	**1.22e-15**	***

Multiple R-Squared: 0.4814,Adjusted R-squared: 0.4761
F-statistic: 90.95 on 1 and 98 DF, p-value: 1.225e-15

Dobimo iste rezultate kot s t-testom za dva neodvisna vzorca in enakimi variancami

t = -9.537, df = 97.82, p-value = 1.245e-15
95% confidence interval:
-17.89687 -11.73162
Ocene vzorca:
povprečje
F: 55.269 M: 70.082

Kodiranje

M-> 1 Povprečni Weight M= $\alpha + \beta$ *1 = $\alpha+\beta$

F-> 0 Povprečni Weight F = $\alpha + \beta$ *0 = α

Multipla linearna regresija

$Weight=\alpha+\beta*Height+\lambda*Gender+\varepsilon$

Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-109.69177	4.87072	-22.521	< 2e-16 ***
height	1.00060	0.02949	33.936	< 2e-16 ***
genderM	4.44250	0.53178	8.354	4.72e-13 ***

Residual standard error: 2.176 on 97 degrees of freedom

Multiple R-Squared: 0.9597, Adjusted R-squared: 0.9589

F-statistic: 1155 on 2 and 97 DF, p-value: < 2.2e-16

$Weight=\alpha+\beta*Height+\varepsilon$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-131.54354	5.36023	4.54	<2e-16
height	1.14217	0.03148	36.28	<2e-16

Multiple R-Squared: 0.9307, Adjusted R-squared: 0.93

Interpretacija koeficientov

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-109.69	4.870	-22.521	< 0.001
height	1.00	0.029	33.936	< 0.001
Gender (M vs F)	4.44	0.538	8.354	< 0.001

$Weight=\alpha+\beta*Height+\lambda*Gender+\varepsilon$

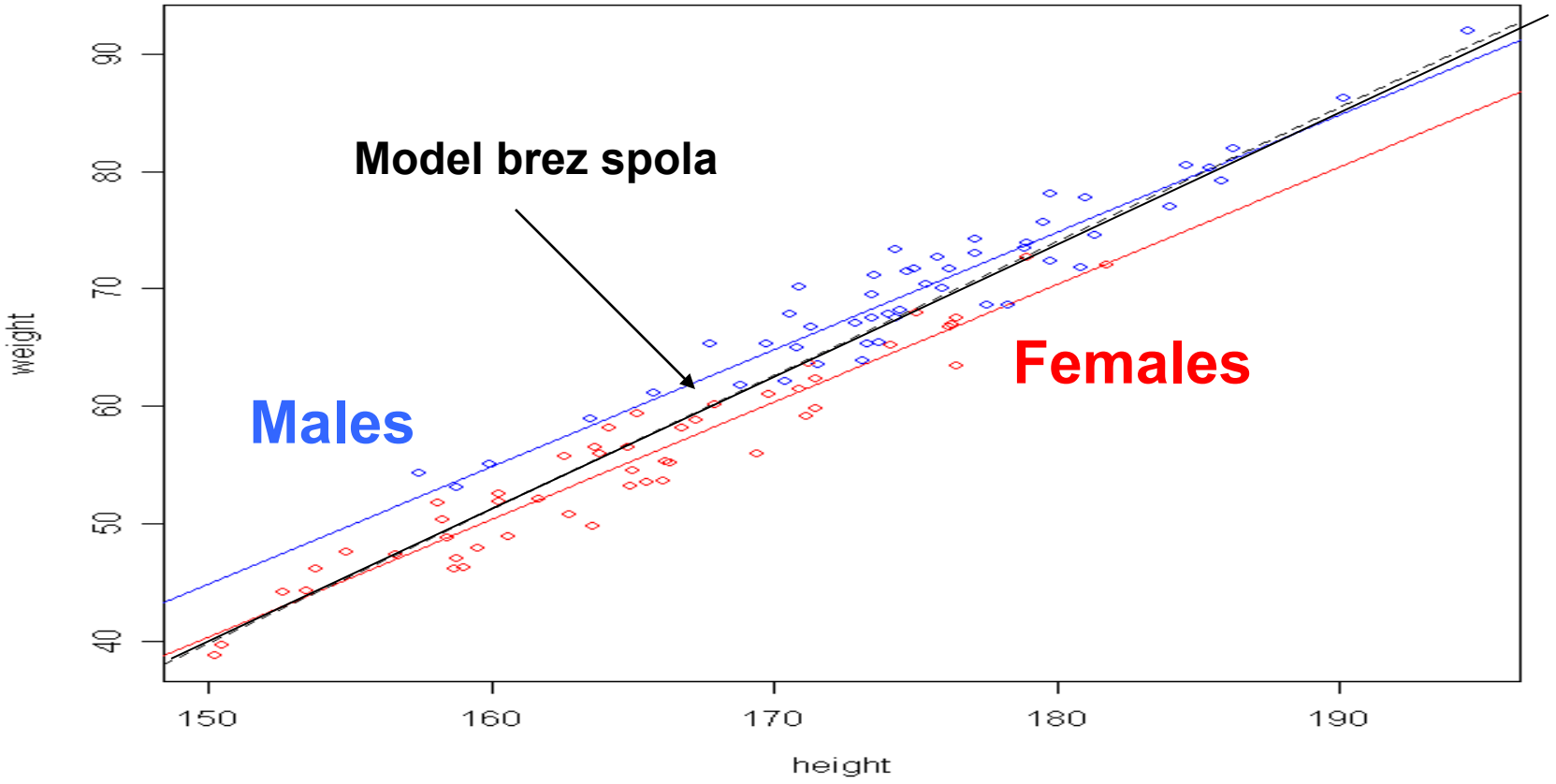
Kodiranje

M-> 1 Povprečni Weight M, če je height=H= $\alpha+\boldsymbol{\lambda*1}+\beta*H$

F-> 0 Povprečni Weight F , če je height=H = $\alpha+\boldsymbol{\lambda*0}+\beta*H$

Razlika povprečne teže med moškimi in ženskami za isto višino

Prileganje the 2 modelov



Multipla linearna regresija

$$Weight=\alpha+\beta*Height+\lambda*Gender+\gamma*Height:Gender+\varepsilon$$

Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-99.30	3.58	-27.75	< 2e-16 ***
height	1.00	0.020	49.00	< 2e-16 ***
genderM	2.93	5.49	0.53	0.595
Height:gender	0.11	0.03	3.45	0.0008

Residual standard error: 0.9654 on 96 degrees of freedom
Multiple R-squared: 0.9945, Adjusted R-squared: 0.9943
F-statistic: 5744 on 3 and 96 DF, p-value: < 2.2e-16

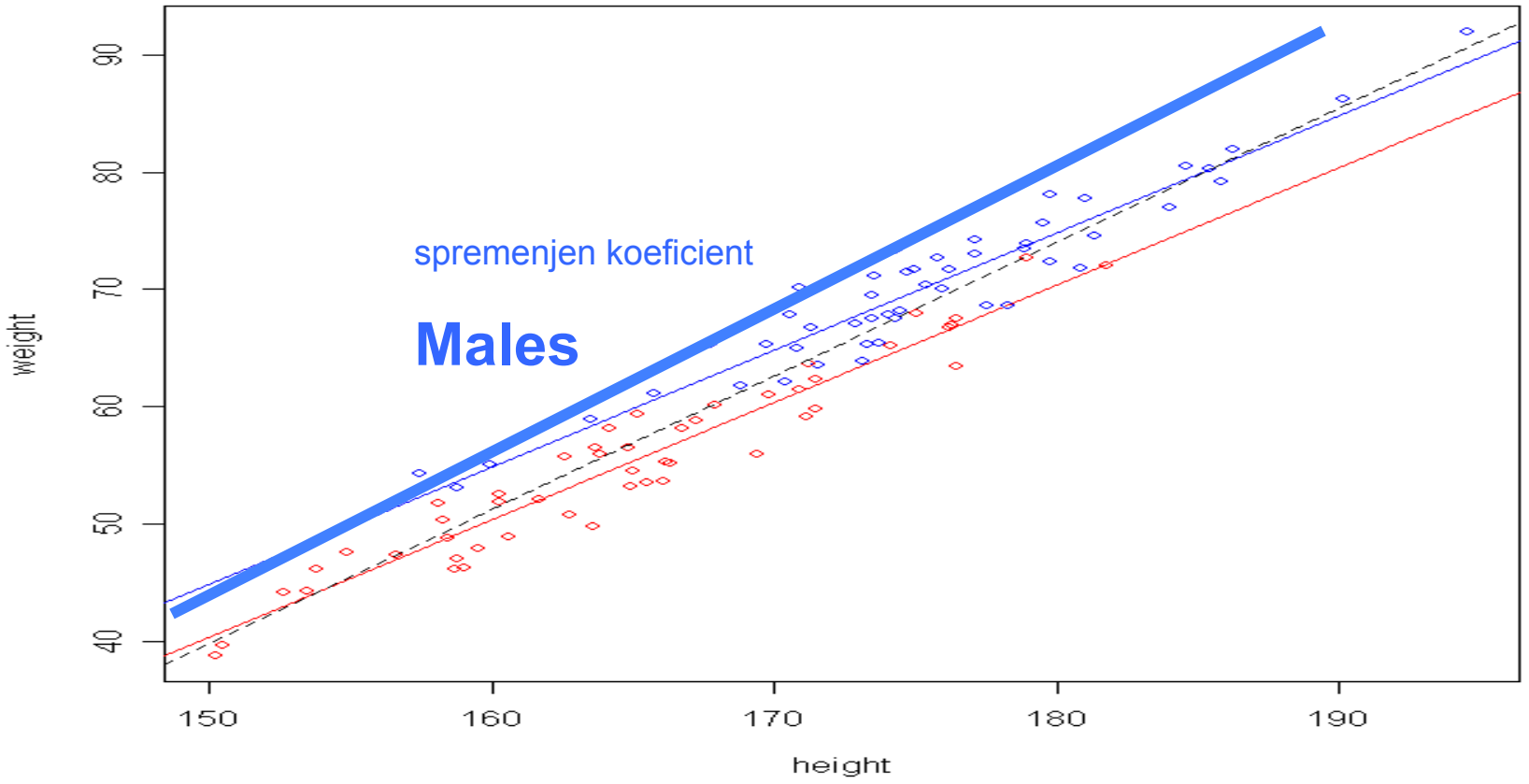
Kodiranje

M-> 1 Povprečni Weight M, če je height=H$\alpha+\lambda*1+\beta*H+\gamma*1*H$

F-> 0 Povprečni Weight F , če je height=H$\alpha+\lambda*0+\beta*$height** $\gamma*0*H$

"spremenjen" koeficient pri height za M

$\lambda+\gamma*H$


spremenjen koeficient
Males
weight
height
40
50
60
70
80
90
150
160
170
180
190

Še en primer

- $y = \alpha + \beta * x_1 + \lambda * x_2 + \varepsilon$

Coefficients:

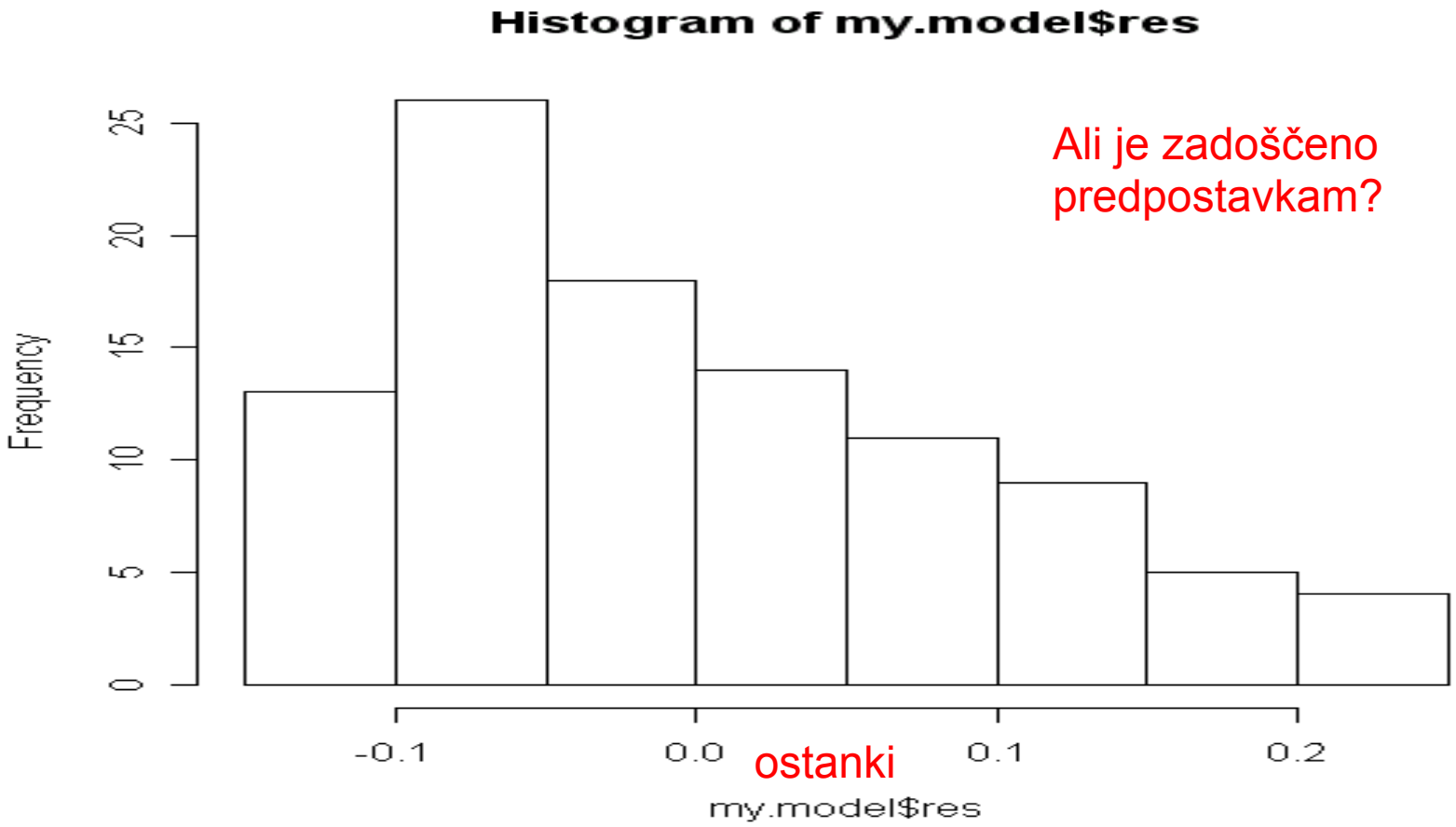
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.42232	0.02815	-15.00	<2e-16 ***
x1	1.08323	0.03667	29.54	<2e-16 ***
x2	1.04614	0.03610	28.98	<2e-16 ***

Residual standard error: 0.09761 on 97 degrees of freedom

Multiple R-Squared: 0.9473, Adjusted R-squared: 0.9462

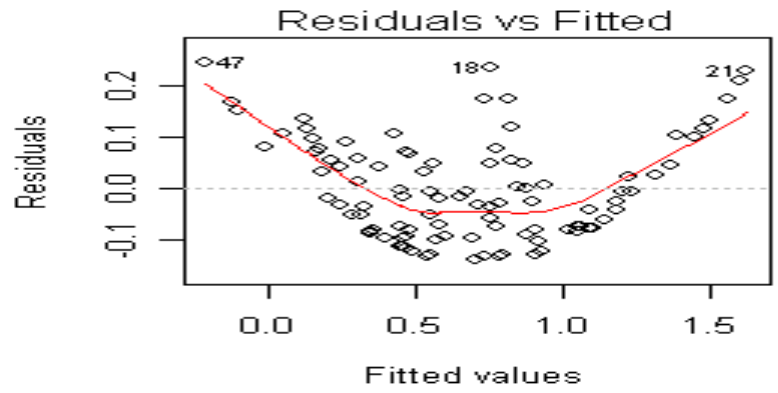
F-statistic: 871.2 on 2 and 97 DF, p-value: < 2.2e-16

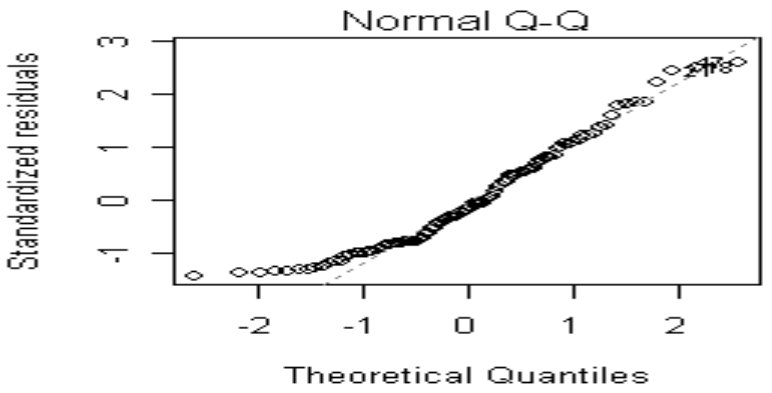
Porazdelitev ostankov

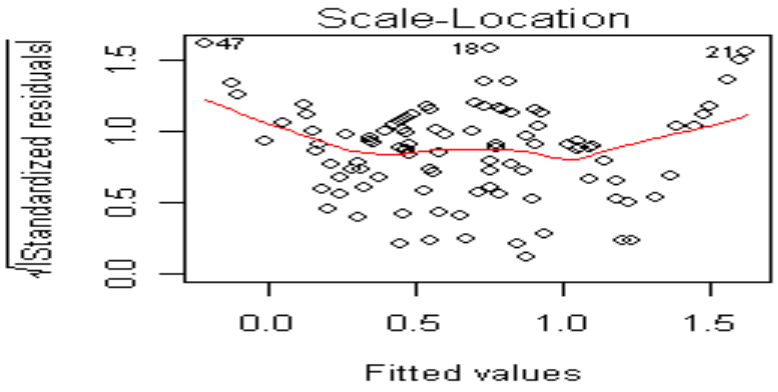


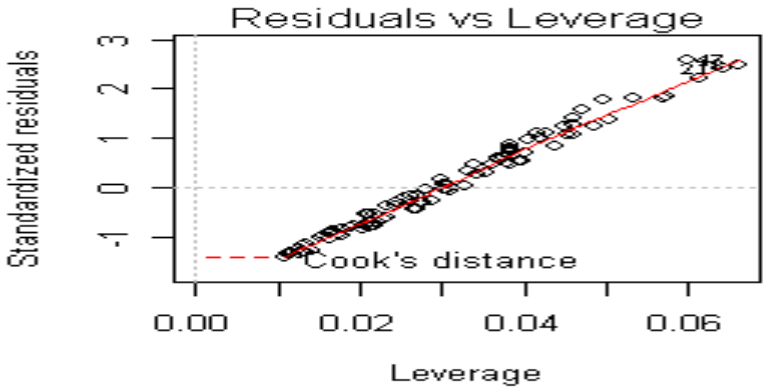
Ali je zadoščeno predpostavkam?

Ali je model dober?

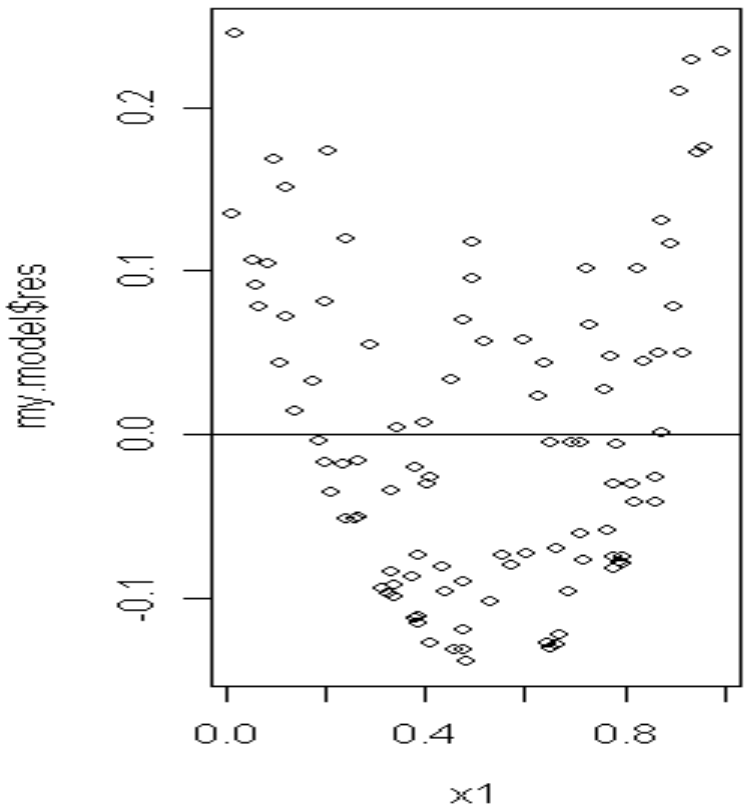


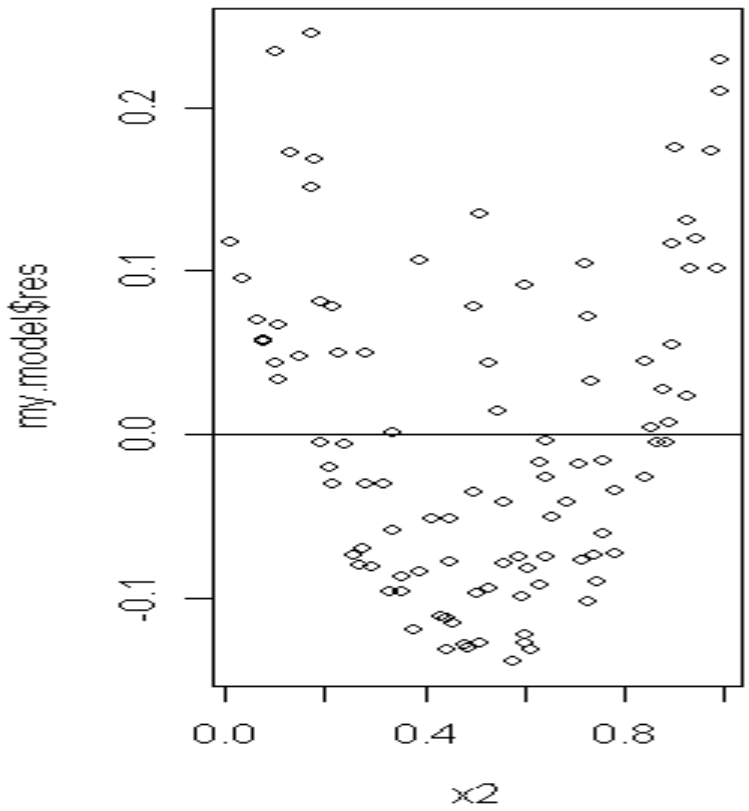






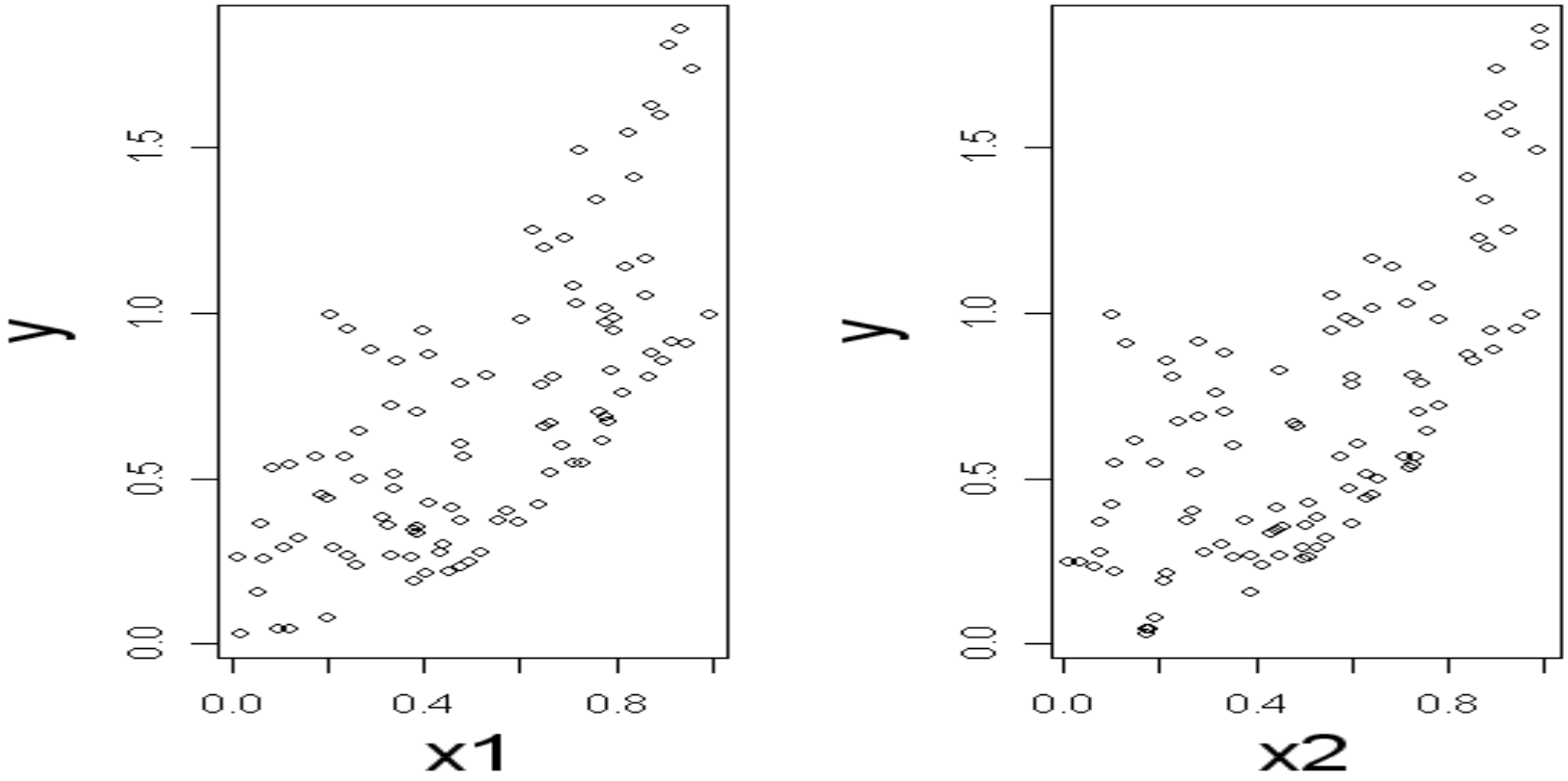
Ali je model dober?





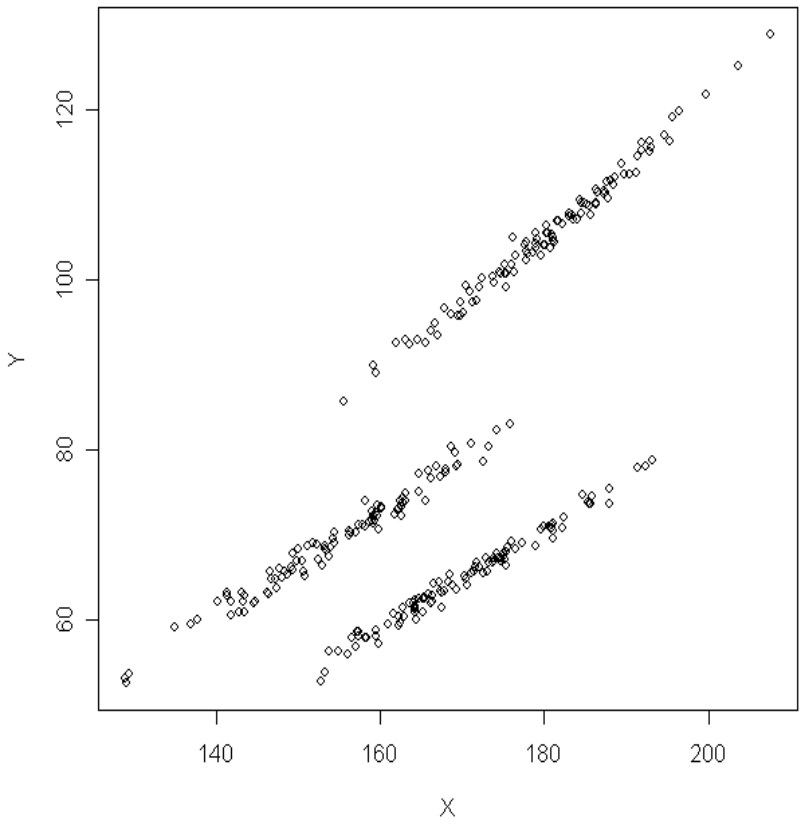
Razsevni diagram primera

(gl. prejšnjo prosojnico)



Multipla linearna regresija

Weight= $\alpha + \beta$ * Height+ λ * Sex



Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-33.512947	1.222303	-27.42	<2e-16	***
X	0.664815	0.007854	84.64	<2e-16	***
sexF	-14.814300	0.224659	-65.94	<2e-16	***
sexM	18.844936	0.275119	68.50	<2e-16	***

Residual standard error: 1.331 on 296 degrees of freedom

Multiple R-Squared: 0.9953, Adjusted R-squared: 0.9952

F-statistic: 2.073e+04 on 3 and 296 DF, p-value: < 2.2e-16

KODIRANJE

"Sex":	β_1	β_2
Children	0	0
Females	1	0
Males	0	1