NAVADNA (BIVARIATNA) LINEARNA REGRESIJA

O regresijski analizi govorimo, kadar želimo opisati povezanost dveh numeričnih spremenljivk. Opravka imamo torej s pari podatkov (x_i, y_i), kjer so x_i vrednosti spremenljivke X, y_i pa vrednosti spremenljivke Y.

Pri tem je treba ločiti med naslednjima možnostima:

1. Na **slučajnem vzorcu** enot izmerimo dve značilnosti (spremenljivki). Nobene od spremenljivk ne kontroliramo. Primer: meritev sistoličnega in diastoličnega krvnega tlaka.
2. Eksperimentalnim enotam določimo vrednost (npr. **dozo**) in merimo nek **izid**. Rezultat so zopet pari vrednosti, vendar je prva spremenljivka, doza, tipično kontrolirana, torej njene vrednosti vnaprej določene. Pogosto več različnih enot dobi enako dozo, se pravi, da imamo več izidov pri isti dozi.

V navadi je naslednja terminologija

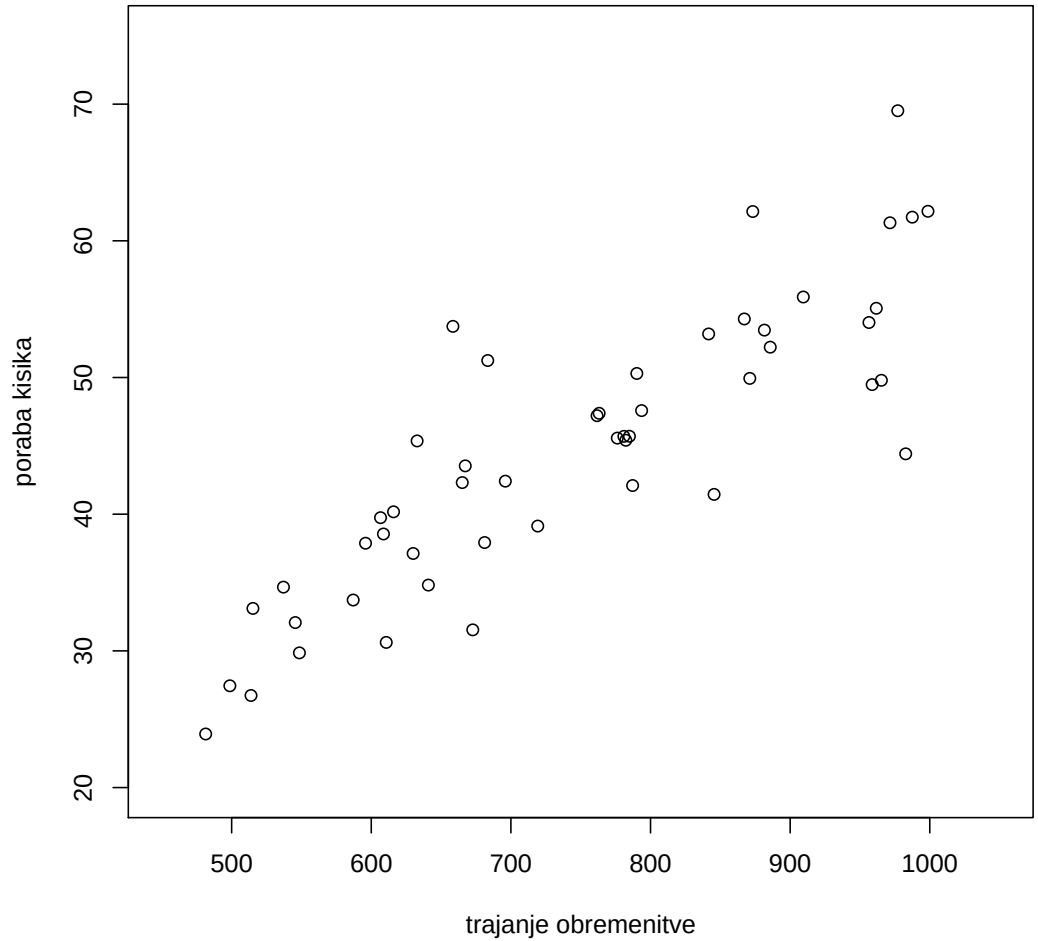
$$X = \text{neodvisna spremenljivka (napovedni dejavnik)}$$
$$Y = \text{odvisna spremenljivka (izid)}$$

Razsevni diagram

Kadar nas zanima povezanost dveh numeričnih spremenljivk, nam prvi vtis o povezanosti dà diagram, v katerem pare vrednosti narišemo kot točke v koordinatnem sistemu.

Primer: Obremenitveni test

Na obremenitvenem testu so 50 moškim izmerili **trajanje obremenitve** in **porabo kisika**. Porabo kisika izražamo v mililitrih na kilogram telesne mase na minuto. Na spodnjem razsevnem diagramu vidimo, da poraba kisika v povprečju narašča s trajanjem obremenitve in to nekako linearno.



V nadaljevanju se bomo omejili na linearno povezanost, kar se morda zdi velika omejitev, a bomo pozneje pokazali, da temu ni tako.

Cilja naše analize bosta dva:

1. Kvantitativni opis in evalvacija povezanosti med spremenljivkama.
2. Napoved Y, če je znana vrednost X.

In ker se bomo omejili na primer, ko je povprečen izid linearno odvisen od napovednega dejavnika, bomo takšni analizi bomo rekli **linearna regresija** (izraz regresija bomo pojasnili kasneje).

Statistični model povezanosti med odvisno in neodvisno spremenljivko bo torej

$$Y = \alpha + \beta X + \epsilon,$$

kjer 'napaka' ϵ predstavlja slučajno variiranje okrog premice.

Od napak bomo zahtevali, da so neodvisne in normalno porazdeljene s povprečjem 0, torej

$$\epsilon \sim \mathcal{N}(0, \sigma^2).$$

Drugače povedano, model predpostavlja, da je pri dani vrednosti x izid Y normalno porazdeljen s (pogojnim) povprečjem

$$E(Y|x) = \alpha + \beta x$$

in (pogojno) varianco

$$Var(Y|x) = \sigma^2.$$

Ker varianca ni odvisna od vrednosti x, smo torej zahtevali, da je variabilnost okrog premice povsod enaka. Temu pogoju pravimo **homoscedastičnost**. Če pogoj ni izpolnjen, govorimo o heteroscedastičnosti.

Naš statistični model ima torej **štiri predpostavke**. Zapišimo jih še enkrat:

1. Opazovanja, se pravi pari meritev, so **neodvisna**.
2. Regresijska funkcija $E(Y|x)$ je **linearna**.
3. Vrednosti variirajo okrog regresijske premice s **konstantno varianco** (homoscedastičnost).
4. Vrednosti Y so okrog regresijske premice **normalno porazdeljene**.

Ustreznost teh predpostavk moramo vedno **preveriti**.

OCENJEVANJE PARAMETROV

Linearni regresijski model ima tri neznane parametre: presečišče α, naklonski koeficient β in varianco σ^2. Prvi problem, ki ga moramo rešiti, ko smo zbrali podatki, je:

Kako izbrati α in β, oziroma katera premica skozi dane točke je najboljša?

Kriterijev za 'najboljšo' premico je več, najpogosteje pa se uporablja tale:

Pri vseh *x* si ogledamo razlike med napovedanimi in izmerjenimi vrednostmi. **Vsota kvadratov** teh razlik naj bo minimalna.

razdelek z zvezdico: največje verjetje

Napovedane vrednosti ležijo na regresijski premici, pri danih x_i jih torej dobimo takole

$$\hat{y}_i = \alpha + \beta x_i.$$

V statistiki je navada, da ocenjene vrednosti označimo s strešico. Seveda je vsak $\hat{y}_i$ ocena za pričakovano vrednost Y pri danem x_i.

Minimizirati moramo torej vsoto

$$SS(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

Pri tem *SS* pomeni sum-of-squares (vsota kvadratov) in čeprav gre za okrajšavo za angleški izraz, je zaradi njene pogostosti ne bomo nadomeščali s slovensko.

Opravka imamo torej s funkcijo dveh spremenljivk (α in β), ki jo z nekaj znanja matematike zlahka minimiziramo. Postopka tukaj ne bomo opisovali, navedimo le končni rezultat:

$$\hat{\beta} = \frac{\sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\sum_i (x_i - \bar{x})^2}$$

in

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Ocenjena premica ima torej enačbo

$$y = \hat{\alpha} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x}),$$

od koder razberemo, da gre premica skozi točko $(\bar{x},\bar{y})$.

Ostane nam še ocena variance σ^2. Izmerjene vrednosti y_i se bolj ali manj razlikujejo od $\hat{y}_i$, razlike

$$r_i = y_i - \hat{y}_i$$

imenujemo **ostanki**. Ostanki so torej **ocenjene napake**. Varianco ostankov ocenimo po formuli

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2} = \frac{SS_{Res}}{n-2},$$

kjer smo z SS_{Res} označili vsoto kvadratov ostankov. Gre seveda za isto vsoto, ki smo jo prej minimizirali, da smo dobili $\hat{\alpha}$ in $\hat{\beta}$, le da tam nismo eksplicitno govorili o ostankih.

Pozoren bralec se bo bržčas vprašal, zakaj smo v formuli za varianco ostankov delili z $n - 2$. Točen odgovor bi bil preveč teoretičen, nek občutek pa dobimo z naslednjim razmislekom: če imamo samo dve točki, lahko skozi njiju potegnemo premico, a ker bosta obe ležali na premici, ne moremo oceniti variance. Potrebujemo torej več točk in šele dodatne točke lahko uporabimo za oceno variance.

Oglejmo si zdaj ostanke.

$$\begin{aligned} r_i = y_i - \hat{y}_i &= y_i - \hat{\alpha} - \hat{\beta} x_i = y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta} x_i \\ &= (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x}) \end{aligned}$$

Kvadrirajmo

$$\begin{aligned} (y_i - \hat{y}_i)^2 &= (y_i - \bar{y})^2 + \hat{\beta}^2 (x_i - \bar{x})^2 - 2\hat{\beta}(x_i - \bar{x})(y_i - \bar{y}) \\ &= (y_i - \bar{y})^2 + \hat{\beta}(x_i - \bar{x})[\hat{\beta} x_i - \hat{\beta}\bar{x} - 2y_i + 2\bar{y}] \\ &= (y_i - \bar{y})^2 + \hat{\beta}(x_i - \bar{x})[\hat{\beta} x_i - \hat{\beta}\bar{x} - 2(\bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_i) + 2\bar{y}] \\ &= (y_i - \bar{y})^2 - \hat{\beta}^2 (x_i - \bar{x})^2 \end{aligned}$$

Če sedaj seštejemo, vidimo, da vsoto kvadriranih ostankov lahko zapišemo takole

$$SS_{Res} = \sum_i (y_i - \bar{y})^2 - \hat{\beta}^2 \sum_i (x_i - \bar{x})^2,$$

kar pomeni, da je vsota kvadratov ostankov enaka vsoti kvadratov odklonov izmerjenih vrednosti od povprečja, **zmanjšani** za vrednost, ki je odvisna od povezanosti med X in Y. K razstavljanju vsote kvadratov odklonov od povprečja se bomo še vrnili, za zdaj opazimo le, da bomo pri močnejši povezanosti (večjem $\hat{\beta}$) odšteli več in bo torej vsota kvadratov ostankov manjša.

Primer: Obremenitveni test (nadaljevanje)

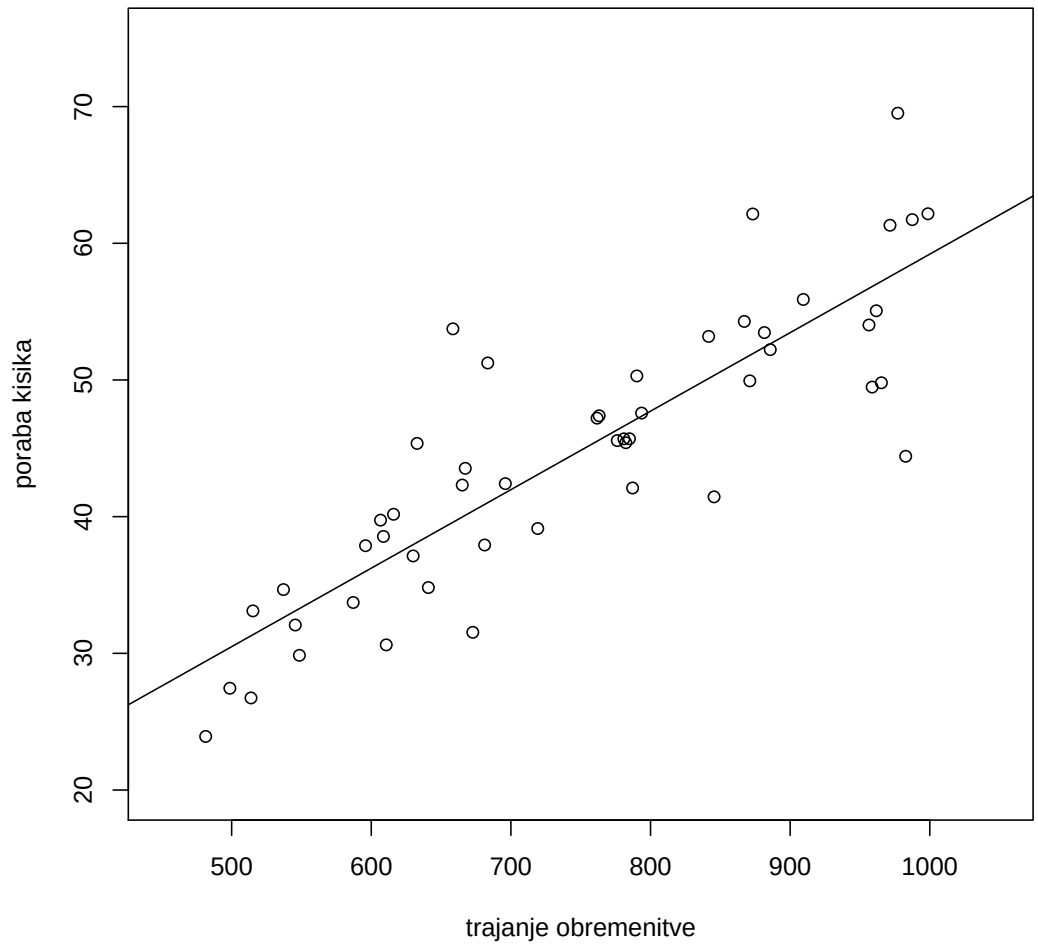
Če sedaj uporabimo formule za $\hat{\alpha}$, $\hat{\beta}$ in $\hat{\sigma}$ na našem primeru podatkov z obremenilnega testa, dobimo

$$\hat{\alpha} = 1{,}765, \quad \hat{\beta} = 0{,}057, \quad \hat{\sigma} = 5{,}348.$$

Ocenjene vrednosti $\hat{y}_i$ bomo iz x_i torej dobili po enačbi

$$\hat{y}_i = 1{,}765 + 0{,}057 \cdot x_i.$$

Ta premica je včrtana na spodnjem grafikonu.


poraba kisika
20
30
40
50
60
70
500
600
700
800
900
1000
trajanje obremenitve

Interpretacija

1. α je seveda vrednost na premici, kadar je x (v našem primeru trajanje obremenitve) enak 0. Takšne vrednosti so redko smiselne, zato koeficientu α ponavadi ne posvečamo posebne pozornosti. Potrebujemo ga pač za to, da pri danem x izračunamo napovedani y.
2. Koeficient β je pri interpretaciji povezanosti dveh spremenljivk najbolj pomemben. Izračunajmo napovedana $\hat{y}$-a pri dveh vrednostih x, ki se med seboj ločita za 1 enoto.

$$\hat{y}(x) = \alpha + \beta \cdot x$$

$$\hat{y}(x+1) = \alpha + \beta \cdot (x+1) = \hat{y}(x) + \beta$$

Torej

$$\hat{y}(x+1) - \hat{y}(x) = \beta.$$

Interpretacija (nadaljevanje)

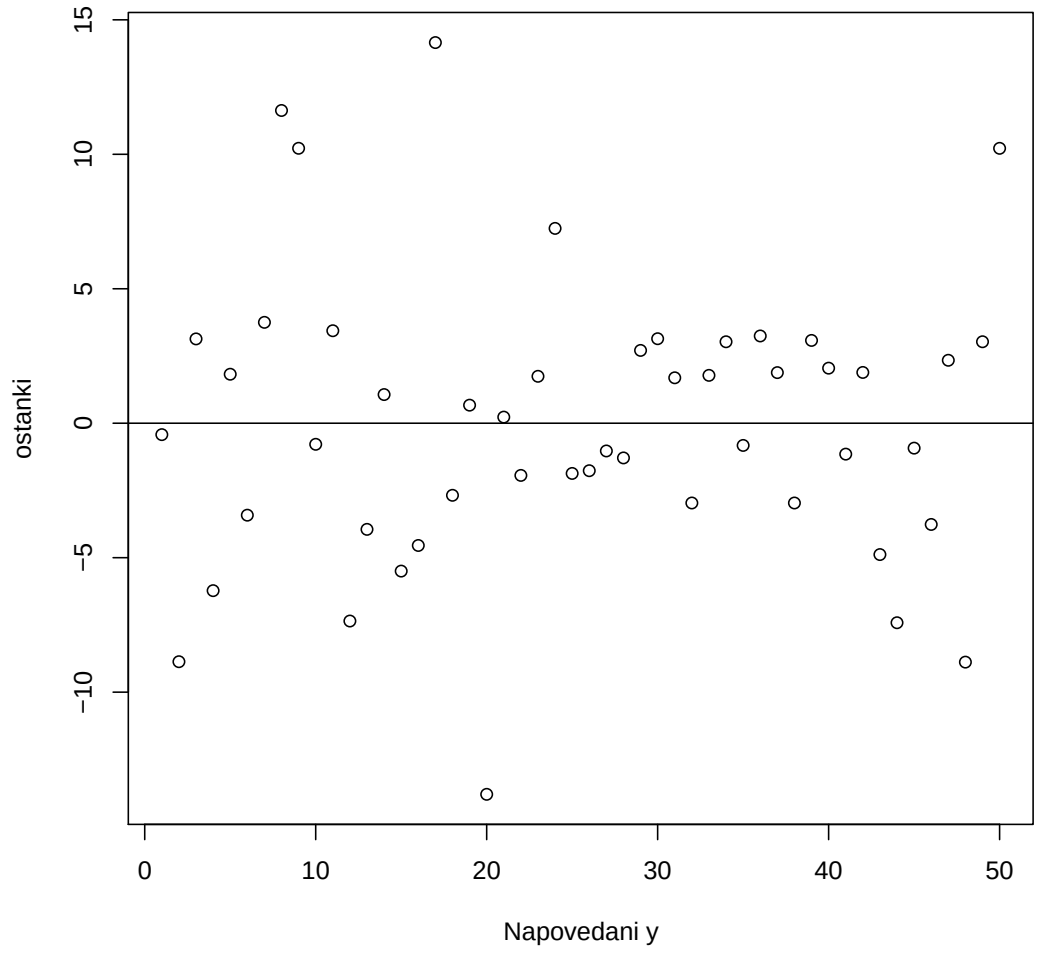
V našem primeru je X (trajanje obremenitve) merjen v sekundah in β potemtakem pomeni spremembo porabe kisika (Y), če se trajanje obremenitve spremeni za 1 sekundo. Ni čudno, da je β majhen. Bolj smiselno bi bilo vedeti, kakšno je povečanje porabe kisika, če je obremenitev daljša za minuto. Seveda je ta vrednost enaka $60 \cdot \beta = 3{,}45$ ml/kg/min.

3. Standardni odklon σ opisuje variabilnost okrog regresijske premice. Ker smo predpostavili normalno porazdelitev, lahko v našem primeru izračunamo, da 95% vseh vrednosti porabe kisika pade v interval $\pm 1{,}96 \cdot 5{,}348 = \pm 10{,}48$ ml/kg/min okrog vrednosti na premici.

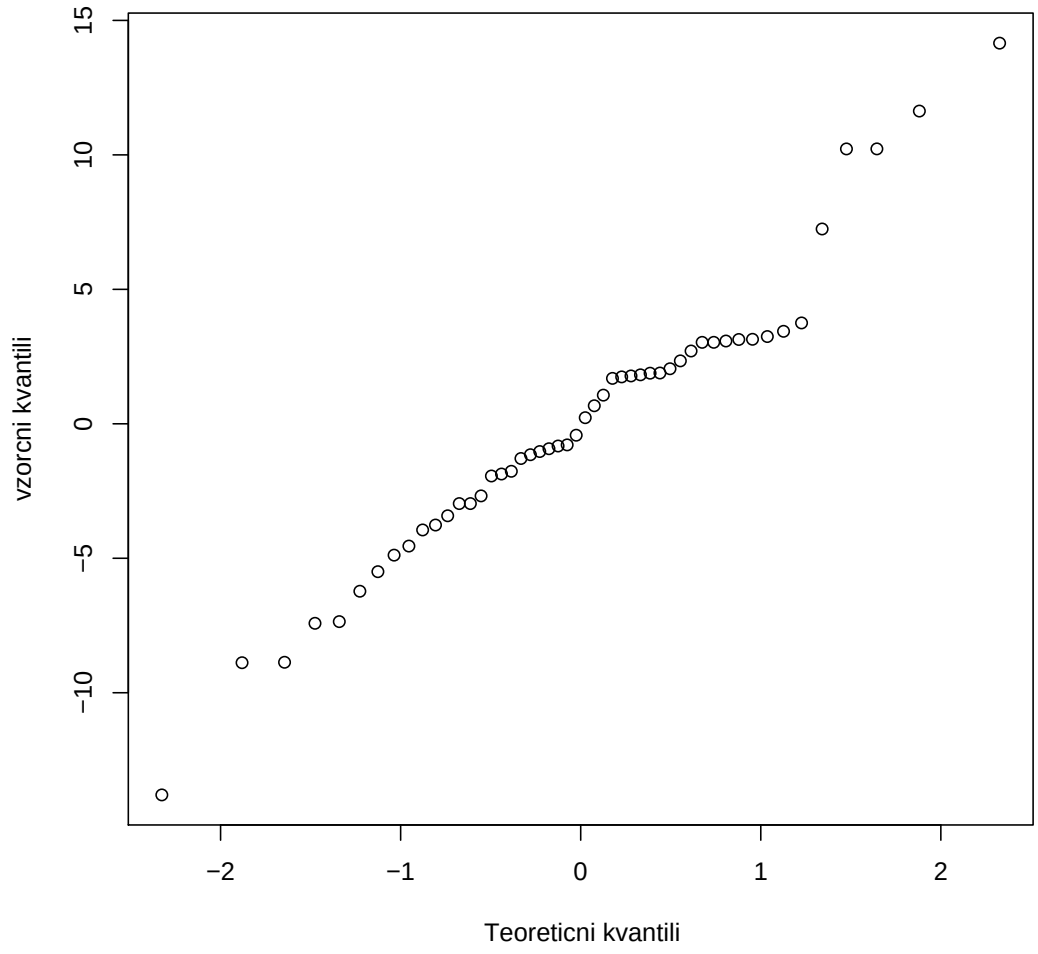
PREVERJANJE PREDPOSTAVK MODELA

1. Če je model pravilen, so **ostanki simetrično razporejeni okrog premice in imajo konstantno varianco**. Graf ostankov r_i glede na izračunane vrednosti $\hat{y}_i$ bi to moral odražati, se pravi, da na njem pričakujemo približno enako razpršenost točk okrog horizontalne črte.
2. Normalnost ostankov lahko preverimo na več načinov, grafično na primer s Q-Q grafom.

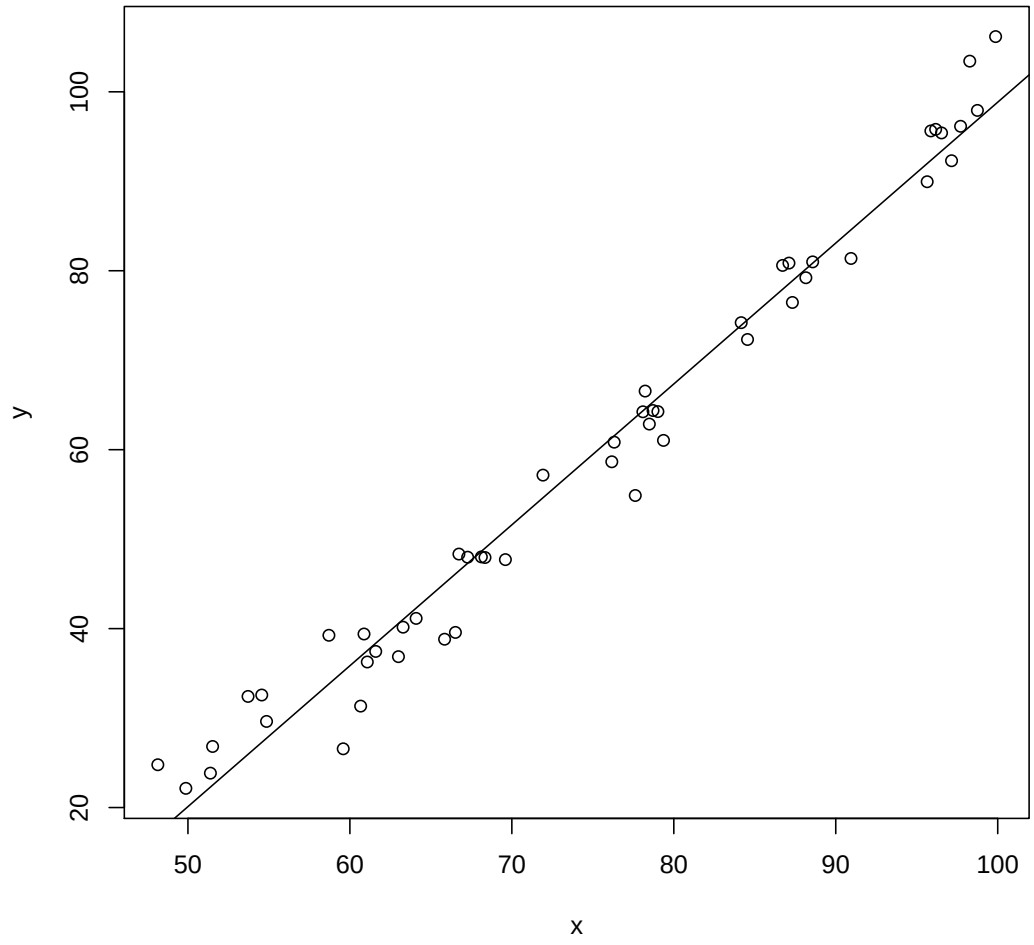
Graf ostankov



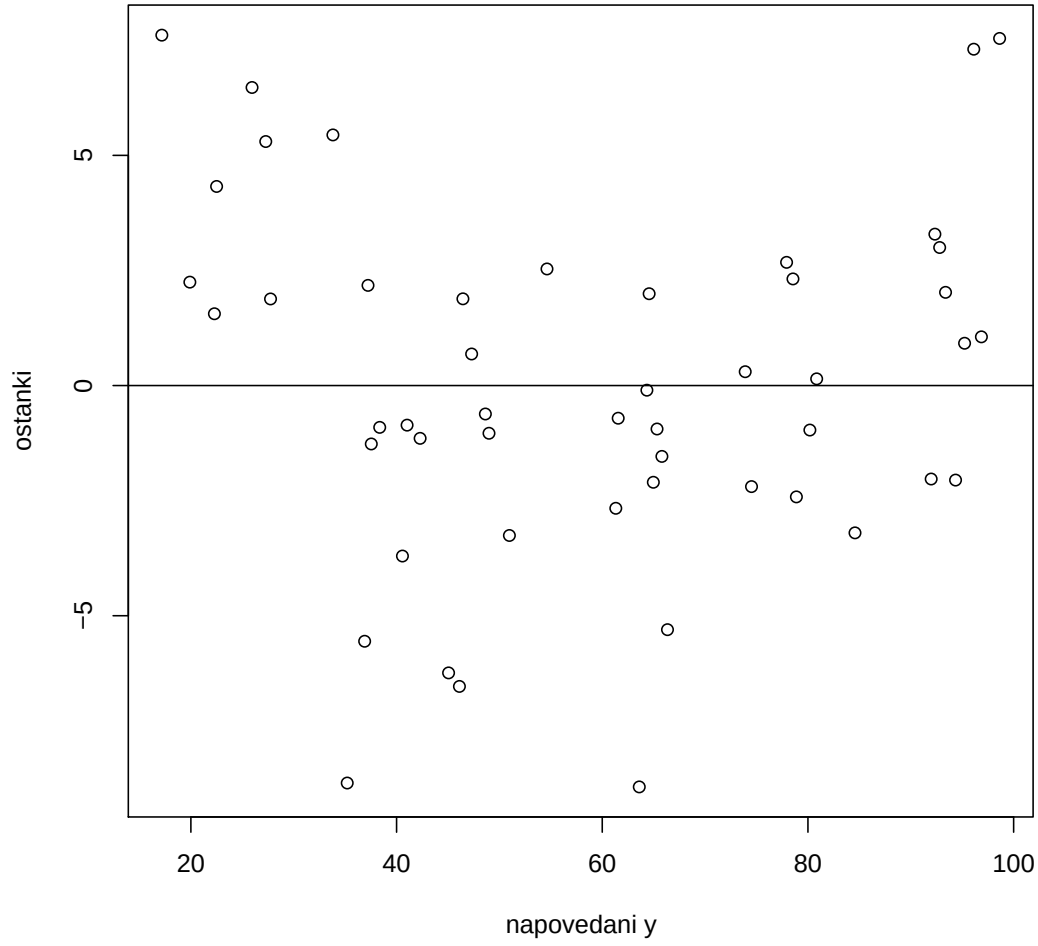
Q-Q graf

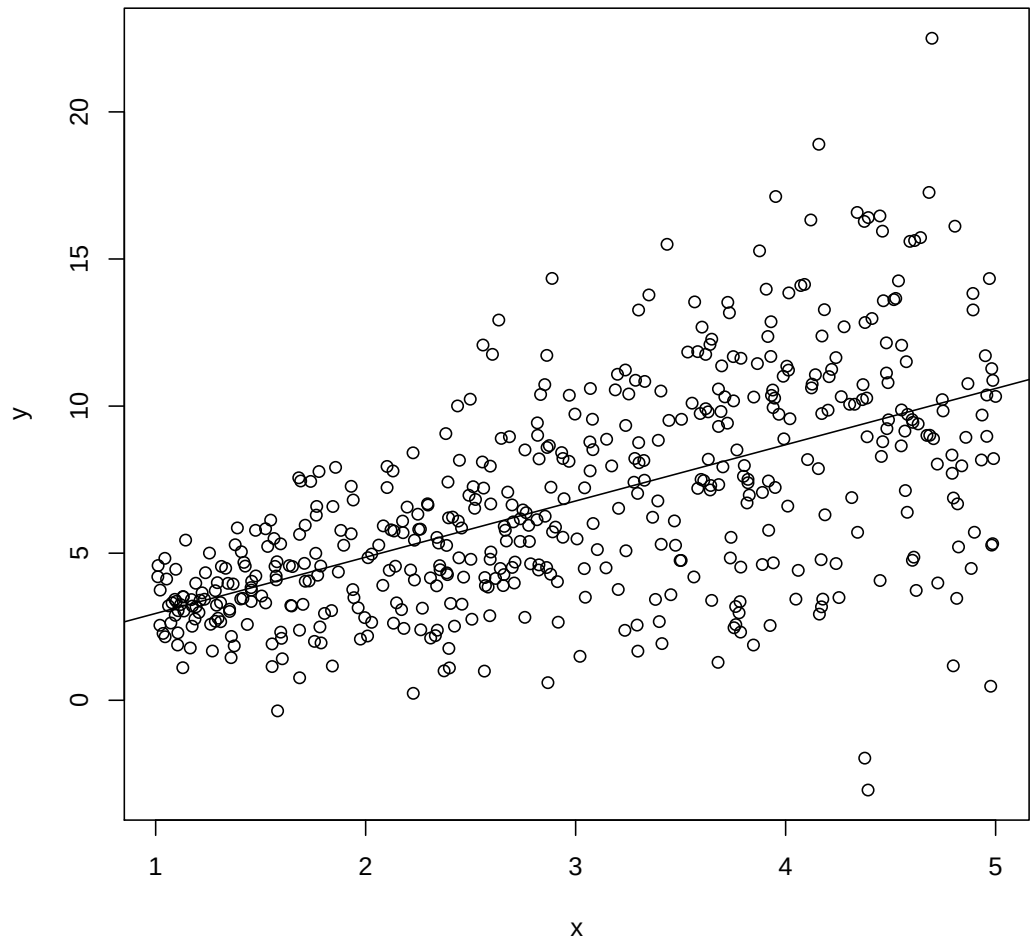


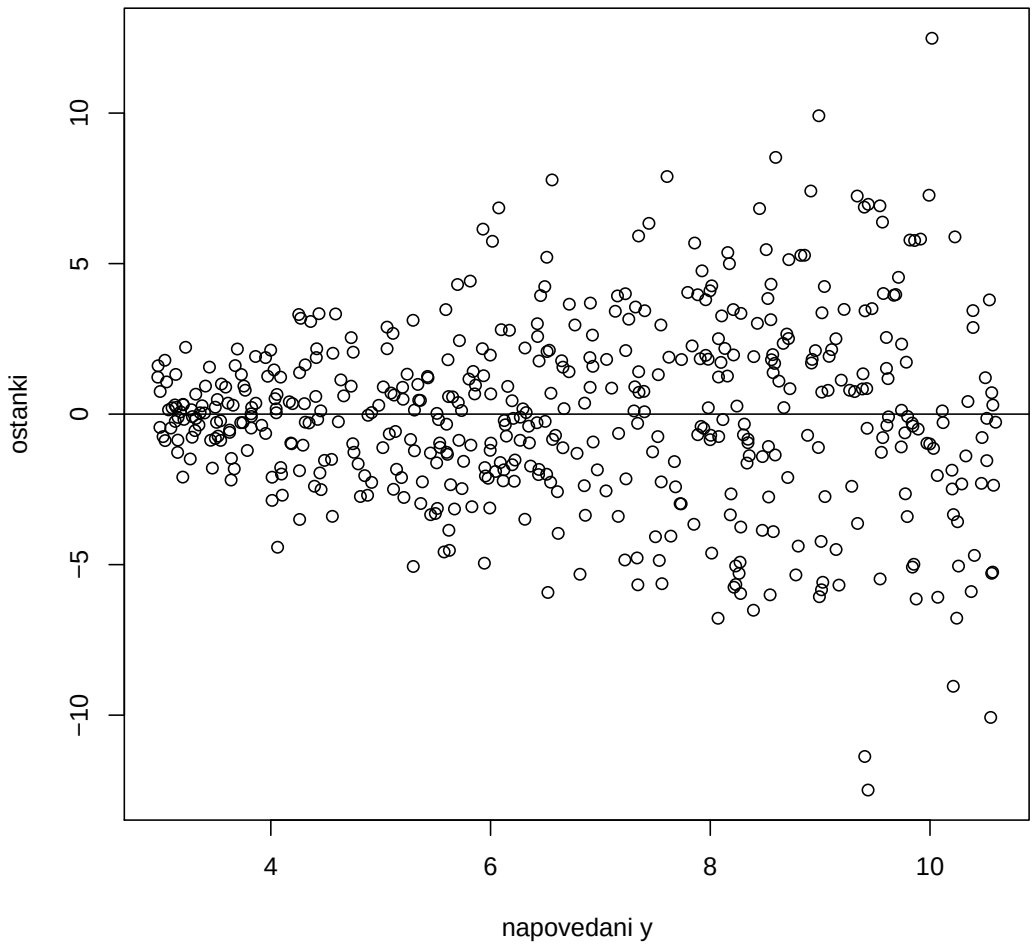
Nelinearna povezanost



Nelinearna povezanost - graf ostankov




y
x
0
5
10
15
20
1
2
3
4
5


ostanki
napovedani y
10
5
0
−5
−10
4
6
8
10

VZORČNE NAPAKE KOEFICIENTOV V LINEARNI REGRESIJI

Pri različnih vzorcih dobimo seveda različne ocene parametrov in vprašati se moramo, kako te ocene variirajo. Pri tem nas posebej zanima naklonski koeficient, ki je odločilen za naše trditve o povezanosti med X in Y. Preden ocenimo njegovo variabilnost, nekoliko poračunajmo.

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i y_i(x_i - \bar{x}) - \bar{y} \sum_i (x_i - \bar{x}) = \sum_i y_i(x_i - \bar{x}),$$

ker je $\sum_i (x_i - \bar{x}) = 0$.

Sedaj privzemimo, da ponavljamo vzorčenje tako, da so x_i vedno isti, vrednosti Y pa naključno variirajo. Prejšnji rezultat upoštevajmo v enačbi za regresijski koeficient $\hat{\beta}$ in potem imamo

$$\begin{aligned} var(\hat{\beta}) &= var\left[\frac{\sum_i Y(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x})^2 var(Y)}{\left[\sum_i (x_i - \bar{x})^2\right]^2} \\ &= \frac{var(Y)}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}. \end{aligned}$$

Ker nas α tipično ne zanima, se formuli za variabilnost $\hat{\alpha}$ izognimo.

Pri danih predpostavkah o porazdelitvi ostankov, potem velja

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$$

Testiranje hipoteze o regresijski premici

Kot vidimo iz prejšnje formule, moramo za opis variabilnosti $\hat{\beta}$ poznati σ. In ker tega praviloma ne poznamo, v formulo za variabilnost $\hat{\beta}$ vstavimo oceno $\hat{\sigma}$. Kot smo že navajeni, to žal spremeni našo trditev o porazdelitvi $\hat{\beta}$, iz lepe normalne v manj lepo porazdelitev t z $n-2$ stopinjama prostosti. Test hipoteze

$$H_0 : \beta = \beta_0$$

je potem

$$t_{\beta=\beta_0} = \frac{\hat{\beta} - \beta_0}{\sqrt{var(\hat{\beta})}}, \qquad \text{sp} = n - 2.$$

Daleč najbolj pogosto seveda testiramo hipotezo $\beta = 0$.

Primer: Obremenitveni test-nadaljevanje

Za ničelno hipotezo $H_0 : \beta = 0$ dobimo $t = 11{,}593$ pri 48-ih stopinjah prostosti, vrednost p pa je $1{,}613 \cdot 10^{-16}$. Ničelno hipotezo torej zlahka zavrnemo.

Vaja:

Kaj se zgodi, če v regresijski analizi namesto neodvisne spremenljivke X želimo uporabiti $Z = X - x_0$, kjer je x_0 neka konstanta?

In kaj, če namesto X uporabimo $U = CX$, kjer je C spet neka konstanta?

RAZSTAVLJANJE CELOTNE VARIABILNOSTI

Recimo, da smo izmerili izid Y in neodvisno spremenljivko X na n posameznikih (enotah). **Celotno variabilnost** izida lahko opišemo z naslednjo vsoto

$$SS_{cel} = \sum_i (y_i - \bar{y})^2,$$

kjer smo z SS označili vsoto kvadratov (**S**um of **S**quares). Ta vsota seveda predstavlja variabilnost, ki je posledica tako biološke variabilnosti, kot tudi variabilnosti zaradi dejstva, da se posamezniki razlikujejo v vrednostih spremenljivke X.

Potem se lahko vprašamo:

Kolikšen del variabilnosti Y gre pripisati variabilnosti X?

Oziroma

Kolikšno variabilnost bi videli, če bi vsi posamezniki imeli enako vrednost X?

Vemo že, da velja

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \bar{y})^2 - \hat{\beta}^2 \sum_i (x_i - \bar{x})^2$$

oziroma

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{\beta} x_i - \hat{\beta}\bar{x})^2.$$

In ker je

$$\hat{y}_i = \hat{\beta} x_i + \hat{\alpha}$$

ter

$$\bar{y} = \hat{\beta}\bar{x} + \hat{\alpha}$$

lahko drugi člen na desni zapišemo tudi kot $\sum_i (\hat{y}_i - \bar{y})^2$ in imamo

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

To ponavadi zapišemo takole

$$SS_{cel} = SS_{ost} + SS_{reg},$$

kar pomeni, da smo celotno variabilnost razstavili na variabilnost zaradi regresije in variabilnost ostankov. Dokaj očitno je, da bo kvocient

$$\frac{SS_{reg}}{SS_{ost}}$$

majhen, če velja **ničelna hipoteza**, da je $\beta = 0$.

Kaj je še majhno in kaj potem veliko (in v nasprotju z ničelno hipotezo) je seveda vprašanje, na katero pa nam odgovori statistična teorija. Celotna vsota kvadratov ima $n - 1$ stopinj prostosti (to vemo iz vzorčne formule za varianco). Vsota kvadratov ostankov ima $n - 2$ stopinji prostosti, kar tudi že "vemo". Izkaže se (s to teorijo se tu ne moremo ukvarjati), da se tudi stopinje prostosti razdelijo med SS_{ost} in SS_{reg} in sicer takole

$$n - 1 = (n - 2) + 1.$$

Teorija potem pravi, da je kvocient

$$F = \frac{SS_{reg}/1}{SS_{ost}/(n-2)} \tag{1}$$

porazdeljen po porazdelitvi $\mathcal{F}$ z 1 in $n-2$ stopinjami prostosti. Vrednost p je potem

$$\text{vrednost } p = P(\mathcal{F}(1,n-2) \geq F),$$

torej verjetnost, da porazdelitev $\mathcal{F}1,n-2)$ zavzame vrednost, ki je večja od ugotovljene. Kritične bodo torej velike vrednosti kvocienta (1). Z drugimi besedami, ničelno hipotezo $H_0 : \beta = 0$ bomo zavrnili, kadar bo variabilnost zaradi regresije velika v primerjavi z variabilnostjo ostankov.

Rezultate ponavadi uredimo v tabelo, ki ji rečemo tabela analize variance (ANOVA).

Vir	sp	SS	MS	F	Značilnost
Regresija	1	SS_{reg}	$SS_{reg}/1$	F	vrednost p
Ostanki	$n-2$	SS_{ost}	$SS_{ost}/(n-2)$		
Skupaj	$n-1$	SS_{cel}			

Primer: Obremenitveni test

Vir	sp	SS	MS	F	Značilnost
Regresija	1	3843.5	3843.5	134.40	1.613e-15
Ostanki	48	1372.7	28.6		
Skupaj	49	5216.2			

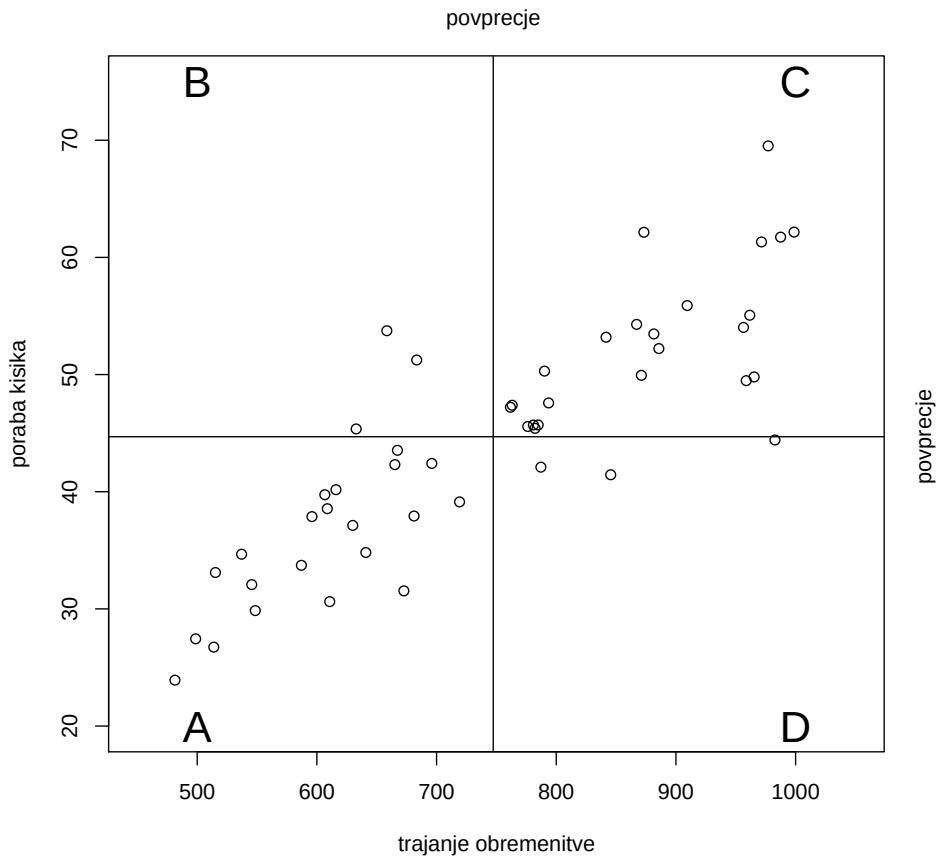
Variabilnost pričakovane vrednosti (pozneje)

KORELACIJA

Oglejmo si količino

$$\sum_i \left[(x_i - \bar{x}) \cdot (y_i - \bar{y}) \right].$$

Zakaj naj bi bilo to zanimivo? Oglejmo si spodnjo sliko.



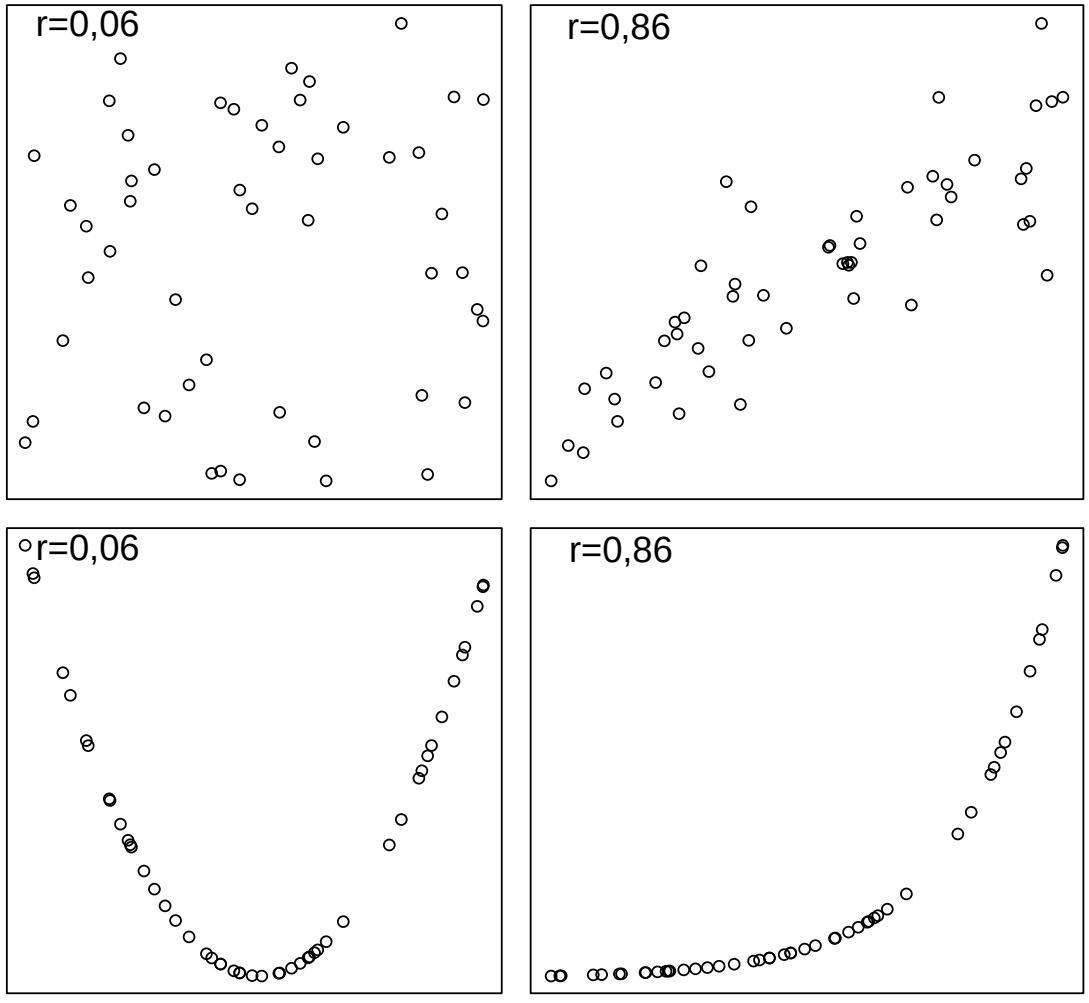
Točke v področjih **A** in **C** bodo prispevale **pozitivne** člene k $\sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})]$, točke v **B** in **D** pa **negativne**. In če so točke skoraj na ravni črti, bo ta vsota velika; pozitivna, če je naklon pozitiven in negativna, če je naklon negativen.

Smiselno je potem definirati empirični **korelacijski koeficient** (**Pearsonov** korelacijski koeficient)

$$r_{XY} = r = \frac{\sum_i [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}},$$

za katerega lahko vidimo, da ima naslednje **lastnosti**:

- Korelacijski koeficient nima enot.
- $-1 \leq r \leq 1$.
- $r_{XY} = r_{YX}$ ter $r_{-XY} = -r_{XY}$.
- Linearna transformacija X ali Y ne spremeni r.
- $r = 1$ ali $r = -1$, kadar vse točke ležijo na premici.



Visok korelacijski koeficient še ne pomeni, da gre za linearen odnos med spremenljivkama! Vedno **narišite podatke!**

PRIČAKOVANA (POPULACIJSKA) KORELACIJA

Korelacijski koeficient smo definirali na vzorcu, vprašati se moramo, kaj je pravzaprav njegova populacijska vrednost. Drugače, če bi zelo velikokrat vzeli vzorec dane velikosti iz populacije, okoli katere vrednosti bi nihal vzorčni korelacijski koeficient.

Naj bosta X in Y dve slučajni spremenljivki in

$$E(X) = \mu_X, \quad E(Y) = \mu_Y,$$

$$Var(X) = \sigma_X^2, \quad Var(Y) = \sigma_Y^2.$$

Kovarianco med X in Y definiramo takole

$$Cov(X, Y) = E\left[(X - \mu_X) \cdot (Y - \mu_Y)\right],$$

pričakovani korelacijski koeficient pa takole

$$\rho = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\cdot\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}.$$

Dvodimenzionalna normalna porazdelitev

Definicija: Par slučajnih spremenljivk (X,Y) je porazdeljen po dvodimenzionalni normalni porazdelitvi, če velja:

1. Vsaka od obeh spremenljivk je normalno porazdeljena, govorimo o robnih normalnih porazdelitvah.
2. Povezanost med obema spremenljivkama je določena s korelacijskim koeficientom ρ_{XY}.
3. X in Y sta neodvisni, če in samo če je $\rho_{XY} = 0$.

Iz gornjih pogojev sledi, da je dvodimenzionalna normalna porazdelitev določena s petimi parametri, in sicer obema povprečjema in standardnima odklonoma

$$(\mu_X, \sigma_X, \mu_Y, \sigma_Y)$$

in korelacijo

$$\rho_{XY}.$$

Zaoišimo gostoto dvorazsežne normalne porazdelitve

$$f(x,y) = K \cdot exp\left[-\frac{z}{2(1-\rho^2)}\right]$$

kjer je

$$K = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

in

$$z = \left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]$$

Trodimenzionalna slika bivariatne normalne.

Zakaj sploh govorimo o dvorazsežni normalni porazdelitvi? Razlog bo razviden iz naslednjega:

Naj bo par (X, Y) porazdeljen po **dvorazsežni normalni porazdelitvi**. Potem velja, da sta **pogojni porazdelitvi** X in Y tudi **normalni** in velja

$$E(Y|X = x) = \mu_{Y|X} = \alpha + \beta \cdot x$$

kjer je

$$\beta = \rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X} \quad \text{in} \quad \alpha = \mu_Y - \beta \cdot \mu_X.$$

Varianca

$$Var(Y|X = x) = \sigma^2_{Y|X} = \sigma^2_Y \cdot (1 - \rho^2_{XY})$$

ni odvisna od vrednosti x.

Opozorimo na dve zanimivosti:

- $1 - \rho_{XY}^2$ je faktor, za katerega se zmanjša varianca Y pri danem x.
- Povezava med regresijskim in korelacijskim koeficientom velja tudi za oceni

$$\hat{\beta} = \hat{\beta}_{Y|X} = r_{XY} \cdot \frac{s_Y}{s_X}$$

Mera pojasnjene variance v linearni regresiji

Videli smo, da v primeru dvorazsežne normalne porazdelitve velja

$$\frac{\text{Varianca pogojne porazdelitve } Y}{\text{Varianca robne porazdelitve } Y} = \frac{\sigma^2_{Y|X}}{\sigma^2_Y} = 1 - \rho^2_{XY}.$$

To pomeni, da je $1 - \rho^2_{XY}$ delež variance Y, ki ga ne moremo pojasniti z variabilnostjo X, oziroma, da je

$$\rho^2_{XY} = \text{delež pojasnjene variance.}$$

Za vzorčno oceno dobimo analogen rezultat

$$\begin{aligned} r_{XY}^2 &= \frac{[\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})]^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} \\ &= \frac{\hat{\beta}^2 \cdot \sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = \frac{SS_{reg}}{SS_{cel}}, \end{aligned}$$

ki velja vedno, torej ni odvisen od predpostavke o normalnosti.

Test hipoteze $\rho_{XY} = 0$

Če je $\rho_{XY} = 0$, potem je testna statistika

$$t = \sqrt{n-2} \cdot \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}}$$

porazdeljena po porazdelitvi t z $n-2$ stopinjami prostosti. Pokažemo lahko, da je ta test t identičen s testom t za ničelni naklon v linearni regresiji.

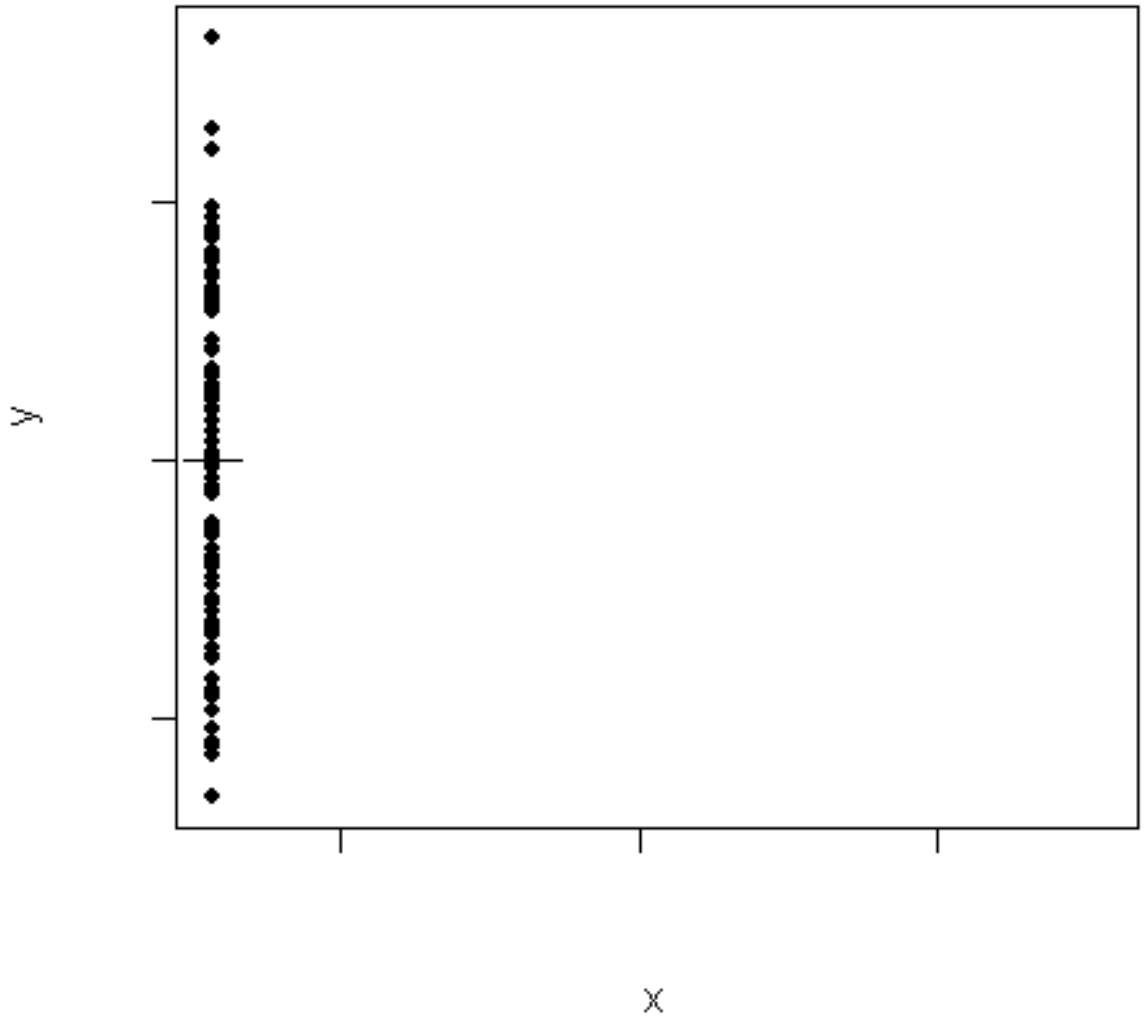
Primer: Obremenitveni test-nadaljevanje

Korelacijski koeficient je 0,86, test t pa 11.593, torej točno toliko, kot pri linearni regresiji.

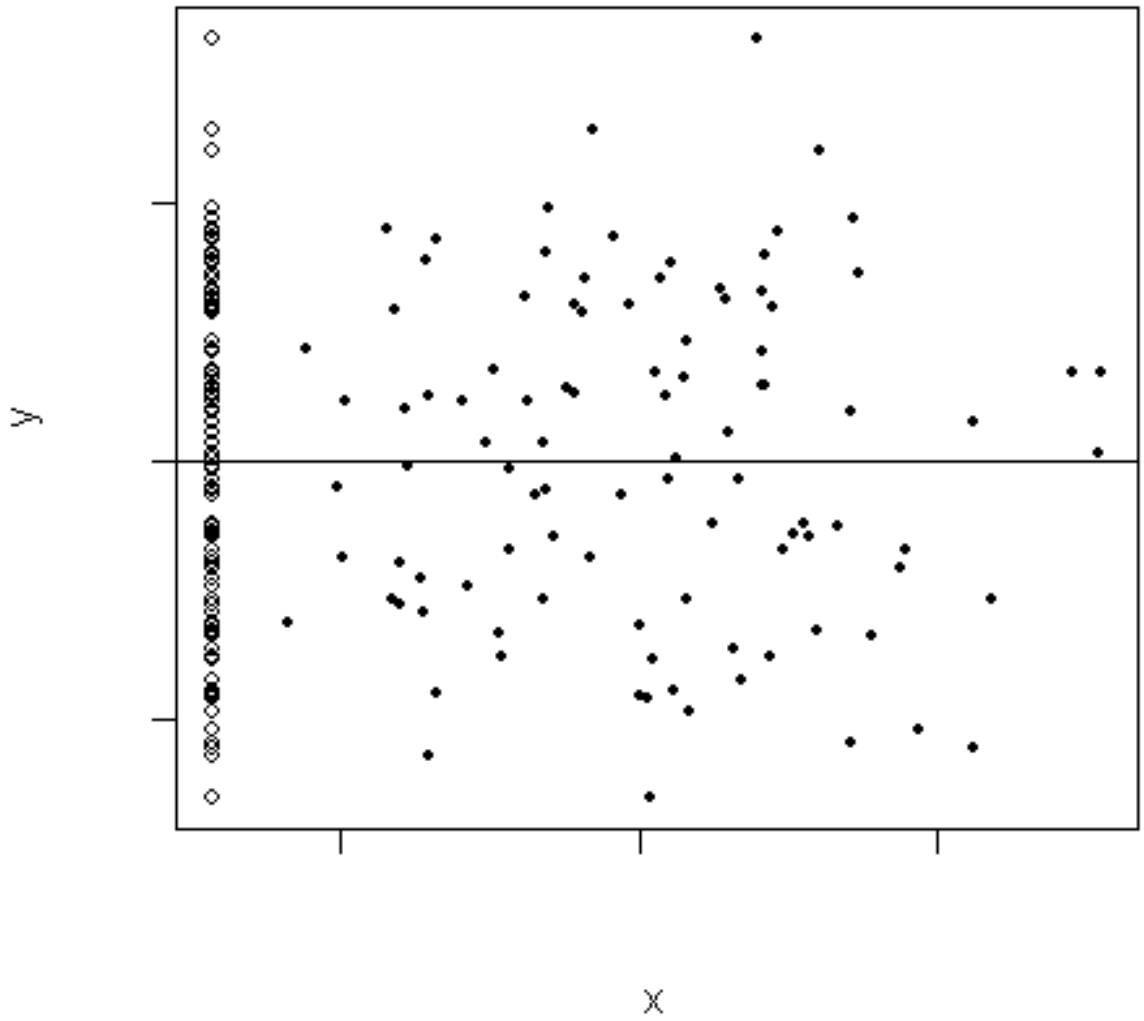
Kvadrat korelacijskega koeficienta je 0,737, kar pomeni, da smo približno 74% variabilnosti porabe kisika pojasnili s trajanjem obremenitve.

Naslednje štiri slike ilustrirajo pomen R^2

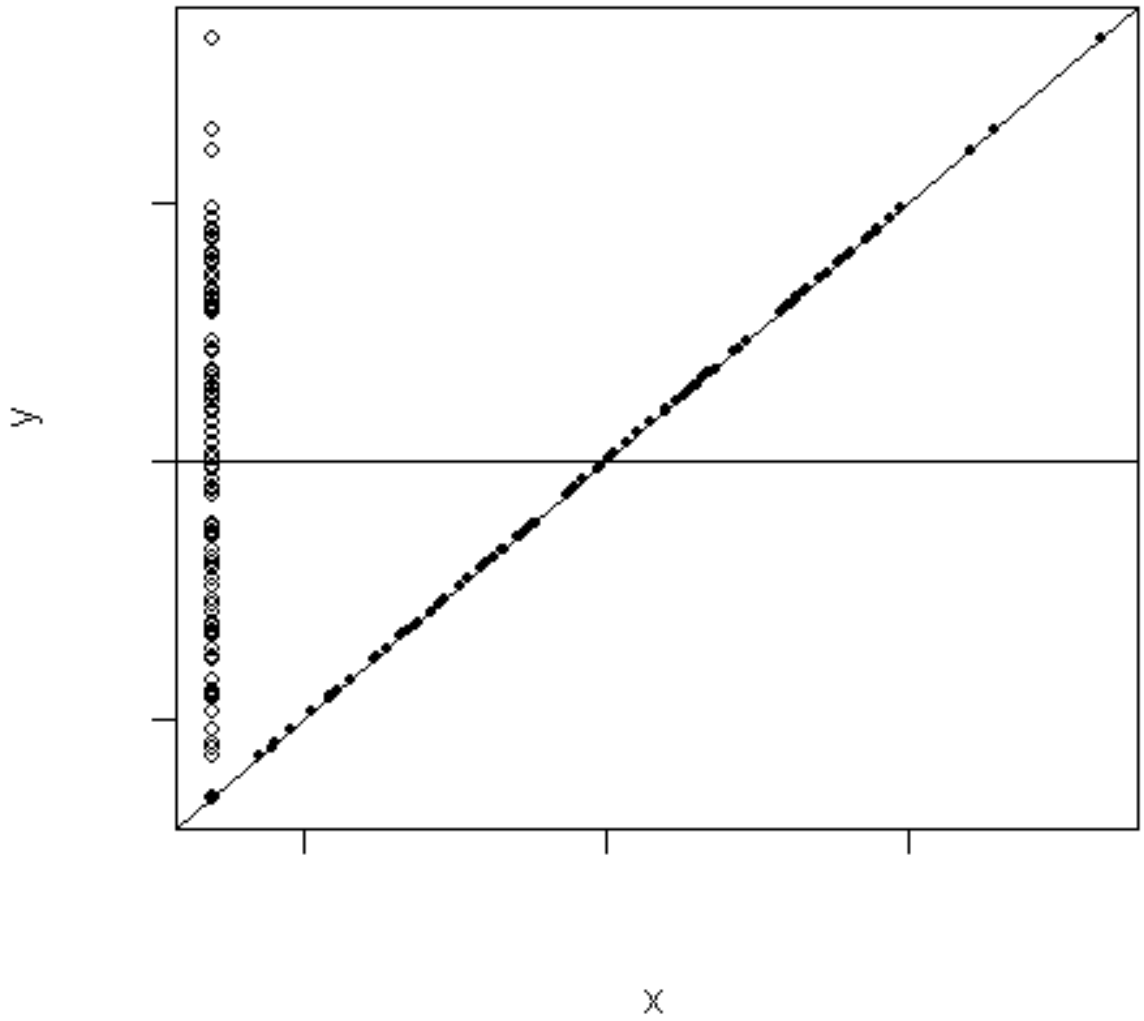
Tole vemo o Y, če ne vemo ničesar o X



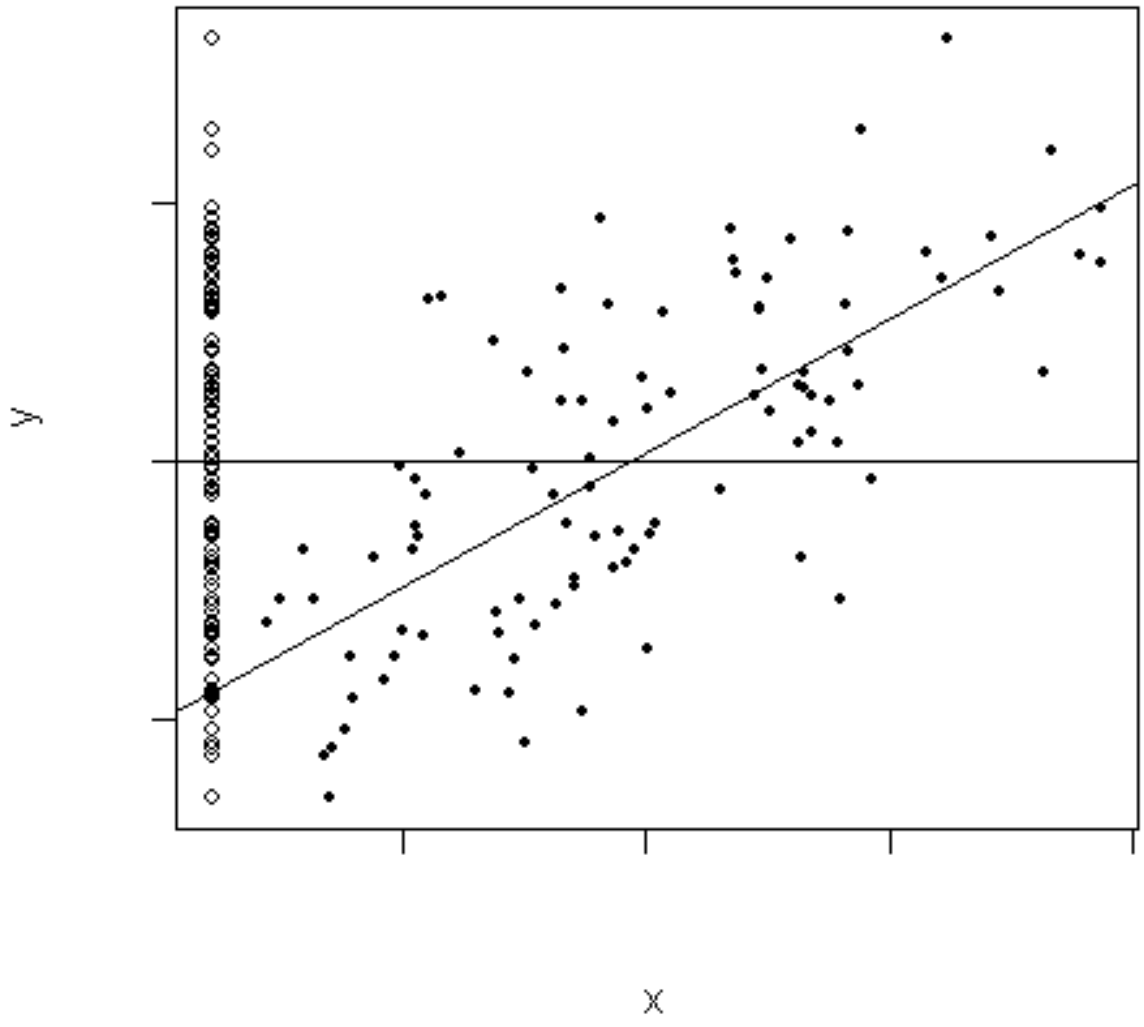
Tole vemo o Y, če poznavanje X ne spremeni ničesar



Tole vemo o Y, če obstaja popolna linearna povezanost z X



In tole vemo o Y, če poznavanje X nekaj pove o variabilnosti Y



MULTIPLA LINEARNA REGRESIJA Pogosto imamo več kot eno neodvisno spremenljivko. Model smiselno razširimo takole

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

kjer 'napaka' ϵ spet predstavlja naključno variiranje okrog hiper ravnine (seveda ne premice).

Vse predpostavke ostanejo enake, metoda ocenjevanja se ne spremeni (razen tega, da moramo rešiti večji sistem enačb).

Nepristranska ocena variance napake je

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n - k - 1} = \frac{SS_{Res}}{n - k - 1}$$

Tudi interpretacija koeficientov je enaka kot v bivariatnem primeru.

Razlika pa nastane pri testiranju hipotez. Sedaj namreč lahko postavimo različne ničelne hipoteze.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_k + \beta_k X_k + \epsilon,$$

nelinearnosti!