Uvod

Statistični model

$$\text{Odziv} = \text{Signal} + \text{Šum}.$$

Odziv	odzivna spremenljivka, odvisna spremenljivka
Signal	sistematična komponenta odzivne spremenljivke
Šum	slučajna komponenta odzivne spremenljivke

S statističnim modelom želimo čim bolje **oceniti signal in šum**.

Uvod

Linearni model (enostavna linearna regresija):

$$y_i = \mu_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- y_i **odzivna spremenljivka**
- ε_i **napaka**, slučajna komponenta odzivne spremenljivke
- μ_i **pričakovana vrednost** y_i, sistematična komponenta odzivne spremenljivke
- x_i **napovedna/neodvisna spremenljivka**
- β_0, β_1 **parametra** linearnega modela

Odzivna spremenljivka y_i je v **linearni zvezi s parametri modela**.

Nelinearni eksponentni model

$$y_i = \beta_0 e^{\beta_1 x_i} + \varepsilon_i,$$

Odzivna spremenljivka y_i

- **ni v linearni zvezi** s parametrom modela β_1
- **je v linearni zvezi** s parametrom β_0 in z napako ε_i.

Statistični model predstavlja redukcijo pogosto obsežnega nabora podatkov na majhno število modelskih parametrov.

- Dober statistični model podatke reducira tako, da lahko na podlagi interpretacije parametrov naredimo smiselne odločitve.
- Model se dobro prilega podatkom, če sistematični del modela dobro opiše variabilnost odzivne spremenljivke, posledično je negotovost majhna.
- Model je dober, če je parsimoničen, kar pomeni, da vsebuje smiselno majhno število parametrov.
- Pri modeliranju je vedno treba narediti kompromis med kompleksnostjo in interpretabilnostjo modela.

Statistično modeliranje je v grobem zaporedje treh korakov, ki jih ciklično ponavljamo dokler diagnostika modela ne pokaže, da je model sprejemljiv:

- začasna formulacija modela
- ocenjevanje parametrov
- diagnostika modela

Za končni model naredimo **obrazložitev rezultatov modeliranja**.

Za **napovedne spremenljivke** velja:

- so lahko številske ali opisne
- so vnaprej izbrane s strani načrtovalca raziskave
- v načrtovanem poskusu izbira vrednosti napovednih spremenljivk vpliva na statistično sklepanje o vplivih napovednih spremenljivk na odzivno spremenljivko in omogoča vzročno-posledično sklepanje
- če vrednosti napovednih spremenljivk niso izbrane vnaprej, v praksi predpostavimo, da so vsaj točne (brez merskih napak).

Namen statističnega modeliranja:

1. razumevanje izbranega procesa opisanega z odzivno spremenljivko in izbranimi napovednimi spremenljivkami, zanimajo nas povezave med napovednimi spremenljivkami in odzivno spremeljivko (*desriptive model*);
2. proučevanje vzročno-posledične zveze med napovednimi spremenlivkami in odzivno spremeljivko, katere napovedne spremenljivke statistično pomembno vplivanjo na proces in kako (*explanatory model*);
3. napovedovanje odzivne spremenljivke na podlagi novih vrednosti napovednih spremenljivk, kjer vrednosti odzivne spremenljivke niso izmerjene/opazovane (*prognostic model*).

`https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf`

Imamo **odzivno spremenljivko** Y in m **napovednih spremenljivk** X_j, $j = 1, ...m$, ki so številske in/ali opisne. m napovednih spremenljivk generira k **regresorjev**, $X_1, ..., X_k$, $k \geq m$.

Predpostavke linearnega modela:

1. Y je številska spremenljivka, njene vrednosti so **medsebojno neodvisne**.
2. Pričakovana vrednost Y pogojno na $X_1, ..., X_k$ je

$$\mathbb{E}(Y|X_1, ..., X_k) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

   β_j, $j = 0, ..., k$ so **parametri modela**. Y je v linearni zvezi s parametri modela.
3. Varianca Y pogojno na $X_1, ..., X_k$, je konstantna

$$Var(Y|X_1, ..., X_k) = \sigma^2 > 0.$$

Iz m napovednih spremenljivk dobimo k regresorjev $X_1, ..., X_k$, $k \geq m$, na različne načine:

- **številsko spremenljivko** v model vključimo direktno kot **en regresor**; včasih je ta spremenljivka predhodno **transformirana** (npr. *log*). V določenih primerih je številska spremenljivka vključena v model z **več regresorji** (npr. polinomska regresija, zlepki);
- za **opisno spremenljivko** z d vrednostmi se v model vključi $d - 1$ regresorjev z vrednostmi 0 in 1 (neme spremenljivke, *dummy variables*);
- dodatne regresorje lahko dobimo z vključitvijo **interakcij med napovednimi spremenljivkami** v modelu.

Linearni model

Modeliramo na podatkih iz vzorca, ki ima n enot

Vrednosti odzivne in napovednih spremenljivk so dobljene **na vzorcu, ki ima n enot**. Linearni model za i-to enoto, $i = 1, ..., n$, zapišemo

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i,$$

ε_i se imenuje **napaka** (*error*), njene lastnosti so:

$$\mathbb{E}(\varepsilon_i) = 0,$$

$$Var(\varepsilon_i) = \sigma^2 \quad \text{ali} \quad Var(\varepsilon_i) = \frac{\sigma^2}{w_i}, \quad i = 1, ..., n,$$

ε_i so medsebojno neodvisni $Cov(\varepsilon_i, \varepsilon_j) = 0$,

w_i, $i = 1, ..., n$ so **znane pozitivne uteži**.

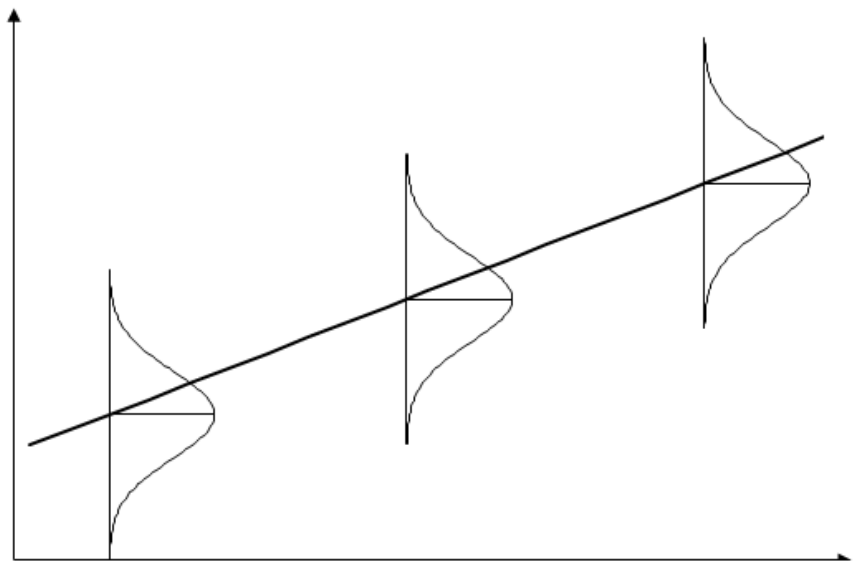
Normalni linearni model: če je lahko predpostavimo, da je porazdelitev Y pri $X_1, ..., X_k$ normalna s povprečjem na regresijski hiper-ravnini in varianco σ^2.

$$Y|X_1, ..., X_k \sim N(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k, \sigma^2),$$

Posledično:

- ε_i so neodvisno enako normalno porazdeljeni $\varepsilon_i \sim iid\ N(0, \sigma^2)$
- $Var(\varepsilon_i) = \sigma^2$ ali $Var(\varepsilon_i) = \frac{\sigma^2}{w_i}$, varianca σ^2 in uteži w_i so konstante

Ilustracija predpostavke linearnega modela na primeru enostavne linearne regresije



Parametre linearnega modela ocenjujemo na podlagi podatkov na vzorcu n enot.
Metode:

- OLS metoda najmanjših kvadratov (*Ordinary Least Squares*)
- WLS tehtana metoda najmanjših kvadratov (*Weighted Least Squares*)
- GLS posplošena metoda najmanjših kvadratov (*Generalised Least Squares*)
- ML metoda največjega verjetja (*Maximum likelihood*)

OLS, minimiramo vsoto kvadratov odklonov y od $\mathbb{E}(y)$:

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2.$$

WLS, minimiramo vsoto tehtanih kvadratov odklonov y od $\mathbb{E}(y)$:

$$S(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} w_i (y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}))^2.$$

Izraz $S(\beta_0, \beta_1, ..., \beta_k)$ parcialno odvajamo po parametrih β_j, $j = 0, ..., k$, in odvode izenačimo z 0. Dobimo **normalni sistem** $k + 1$ **linearnih enačb**. Rešitev tega sistema so **cenilke parametrov**, b_j, $j = 0, ..., k$.

Z modelom **prilagojene vrednosti** (*fitted values*) označimo $\hat{y}_i$:

$$\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_k x_{ik}, \quad i = 1, ..., n.$$

Razliko med dejansko vrednostjo y_i in napovedano vrednostjo $\hat{y}_i$ imenujemo **ostanek** (*residual*) , e_i:

$$e_i = y_i - \hat{y}_i, \quad i = 1, ..., n.$$

Ostanki so e_i nekorelirani z prilagojenimi vrednostmi $\hat{y}_i$, kar uporabljamo pri analizi modela z grafičnimi prikazi.

Varianca ostanka:

$$Var(e_i) = \sigma^2(1 - h_{ii}).$$

h_{ii} se imenuje vzvod (*leverage* ali *hat-value*)

h_{ii} je odvisen od $(x_{i1}, x_{i2}, ..., x_{ik})$, zavzema vrednosti med $1/n$ in 1.

Za izračun $Var(e_i)$ moramo varianco σ^2 oceniti na podlagi podatkov, cenilko variance označimo s^2.

$$\widehat{Var}(e_i) = s^2(1 - h_{ii}).$$

Standardizirani ostanek e_{s_i}:

$$e_{s_i} = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}}, \quad i = 1, ..., n,$$

Če je $n >> k$ velja, da je njihova porazdelitev približno $N(0, 1)$.

Ali so predpostavke modela izpolnjene, ugotavljamo z analizo ostankov in standardiziranih ostankov.

Normalni linearni model

O parametrih modela lahko povemo več

Če lahko privzamemo **normalni linearni model**, poznamo verjetnostne porazdelitve parametrov modela in:

- za vsako cenilko parametra lahko izračunamo njeno standardno napako;
- izračunamo interval zaupanja za vsak parameter modela;
- testiramo lahko statistične domneve o parametrih modela;
- izračunamo napovedi in intervale zaupanja za povprečno napoved in za posamično napoved.

Lastnosti parametrov modela poglejmo najprej na primeru enostavne linearne regresije.

Enostavna linearna regresija

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, ..., n.$$

ε_i so neodvisne slučajne spremenljivke s pričakovano vrednostjo $\mathbb{E}(\varepsilon_i) = 0$ in konstantno varianco $Var(\varepsilon_i) = \sigma^2$ za vsak $i = 1, ..., n$.

β_0, β_1 parametra modela, ki ju bomo ocenili z metodo najmanjših kvadratov OLS.

Izrek 1.1: Po metodi najmanjših kvadratov sta cenilki parametrov β_0 in β_1

$$b_0 = \overline{y} - b_1\overline{x},$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}.$$

SS_{xx} je vsota kvadratov odklonov za x

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i,$$

SS_{xy} je vsota produktov odklonov x in y

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}).$$

Dokaz:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

parcialno odvajamo po β_0 in β_1:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i),$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i).$$

Ko odvoda izenačimo z 0, dobimo sistem dveh linearnih enačb.

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i,$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2.$$

Iz prve enačbe sledi

$$b_0 = \overline{y} - b_1 \overline{x}$$

Ta rezultat uporabimo v drugi enačbi sistema

$$\sum_{i=1}^{n}(x_i y_i - x_i(\overline{y} - b_1\overline{x}) - b_1 x_i^2) = \sum_{i=1}^{n}(x_i y_i - x_i\overline{y} + b_1 x_i\overline{x} - b_1 x_i^2) = 0$$

$$\sum_{i=1}^{n}(x_i y_i - x_i\overline{y}) = b_1 \sum_{i=1}^{n}(x_i^2 - x_i\overline{x}).$$

Iz tega sledi:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i y_i - x_i\overline{y})}{\sum_{i=1}^{n}(x_i^2 - x_i\overline{x})} = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}$$
$$= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{SS_{xy}}{SS_{xx}}$$

Enačba regresijske premice:

$$\hat{y} = b_0 + b_1 x$$

- b_0 presečišče premice z ordinatno osjo
- b_1 naklon premice

Model velja na intervalu $[x_{min}, x_{max}]$.

Ob danih predpostavkah sta cenilki b_0 in b_1 funkciji y_i in posledično tudi ε_i.

Cenilka za varianco napak σ^2 je s^2

Napake: $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ ocenimo z **ostanki**:

$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i,\ i = 1, .., n.$

Definiramo **vsoto kvadratov ostankov** ($SS_{residual}$):

$$SS_{residual} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

Nepristranska cenilka za σ^2:

$$s^2 = \frac{SS_{residual}}{n-2}.$$

V imenovalcu delimo z $n-2$ namesto z n.

Enostavna linearna regresija

Povzetek glede y

Glede odzivne spremenljivke smo do sedaj povedali:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\hat{y}_i = b_0 + b_1 x_i$$

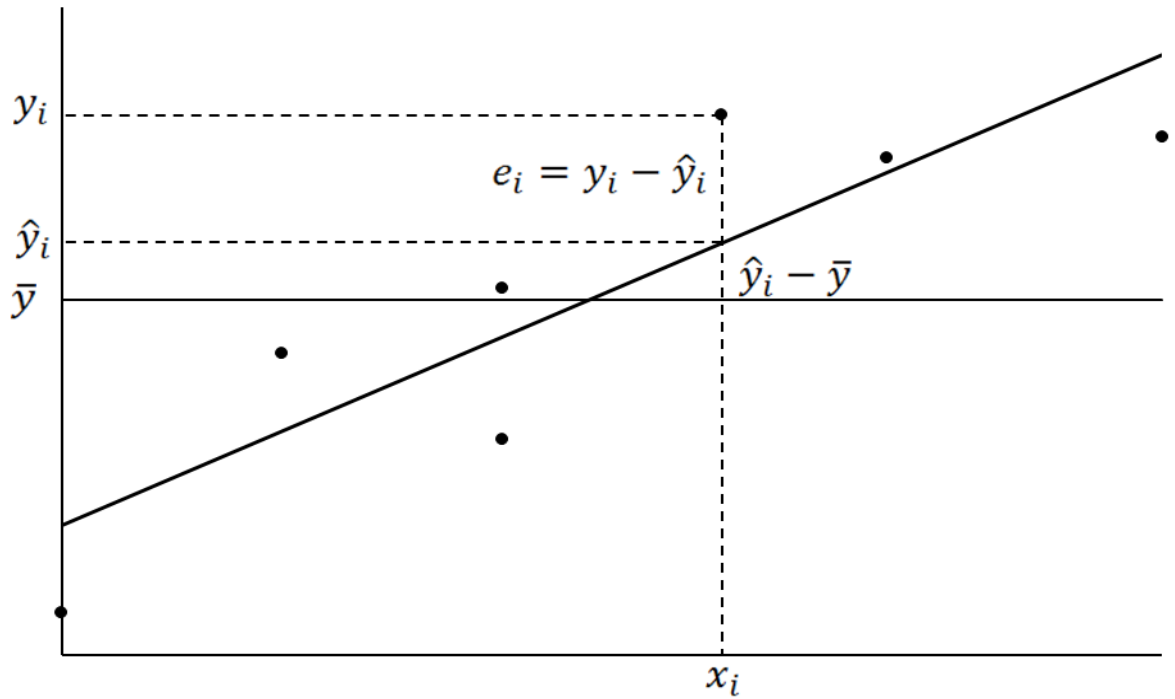
V nadaljevanju bomo videli

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SS_{model} + SS_{residual}$$

Enostavna linearna regresija

$SS_{yy} = SS_{model} + SS_{residual}$

Grafični prikaz, ki je osnova za delitev variabilnosti odzivne spremenljivke na dva dela:



Enostavna linearna regresija

$SS_{yy} = SS_{model} + SS_{residual}$, koeficient determinacije

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

$$SS_{yy} = SS_{model} + SS_{residual}.$$

Koeficent determinacije R^2 je delež variabilnosti za y, ki je pojasnjen z regresijskim modelom:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{SS_{model}}{SS_{yy}}.$$

Koeficient determinacije je enostavna mera za kakovost linearnega regresijskega modela.

Za izračun R^2 ne potrebujemo nobenih predpostavk.

Lastnosti koeficienta determinacije:

- je nenegativna vrednost;
- je manjši ali enak 1; ima vrednost 1, če je $SS_{model} = SS_{yy}$, ko so vse točke na premici;
- R^2 je odvisen od zaloge vrednosti napovedne spremenljivke;
- pri uporabi R^2 moramo biti previdni, saj vsak dodani regresor poveča vrednost R^2, tudi če je vpliv tega regresorja na odzivno spremenljivko statistično nepomemben (multipla regresija).

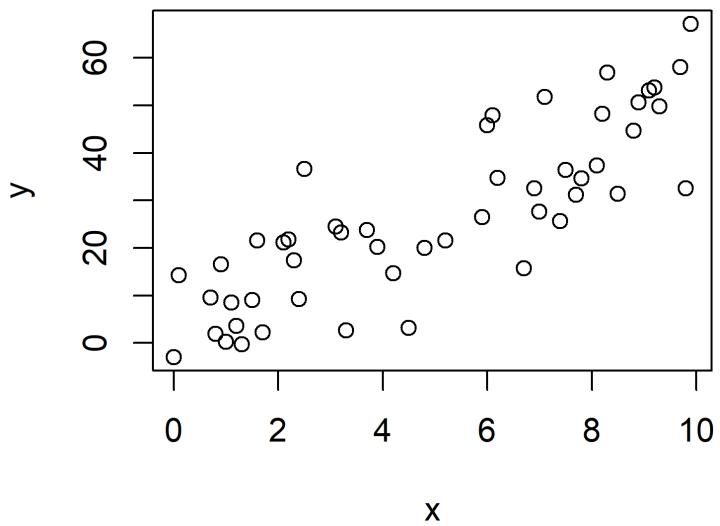
Diagnostika modela je namenjena preverjanju predpostavk linearnega modela. Na podlagi podatkov ocenimo parametre modela in preverimo, ali je bilo tako modeliranje upravičeno. Preverjamo:

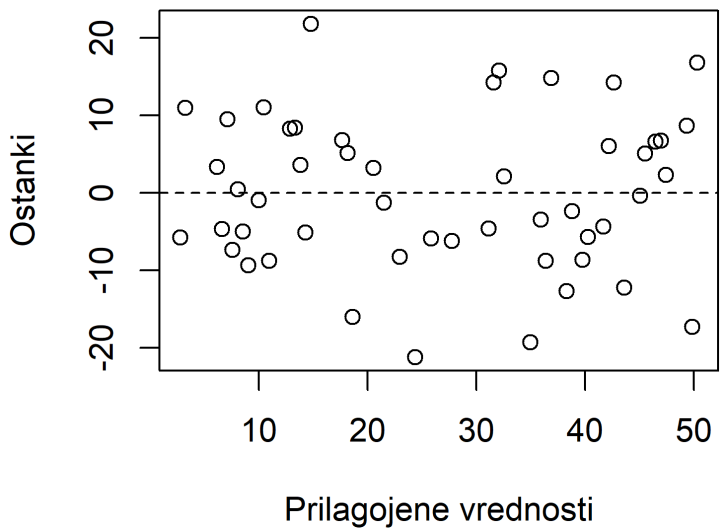
- **linearnost** odvisnosti odzivne spremenljivke od napovedne spremenljivke (razsevni grafikon y glede na x, slika ostankov v odvisnosti od prilagojenih vrednosti, odvsinost ne sme biti vidna);
- **varianca napak** oziroma varianca odzivne spremenljivke pogojno na napovedne spremenljivke **je konstantna** (slika ostankov glede na prilagojene vrednosti);
- **pričakovana vrednost napak je 0** (slika ostankov glede na prilagojene vrednosti);
- **napake so medsebojno neodvisne** (težko preveriti, verjamemo, da so bili podatki pridobljeni z ustreznim načinom vzorčenja, analiza avtokorelacije ostankov).

Enostavna linearna regresija

Diagnostika linearnega modela, primer

Ni odstopanja od predpostavk linearnega modela

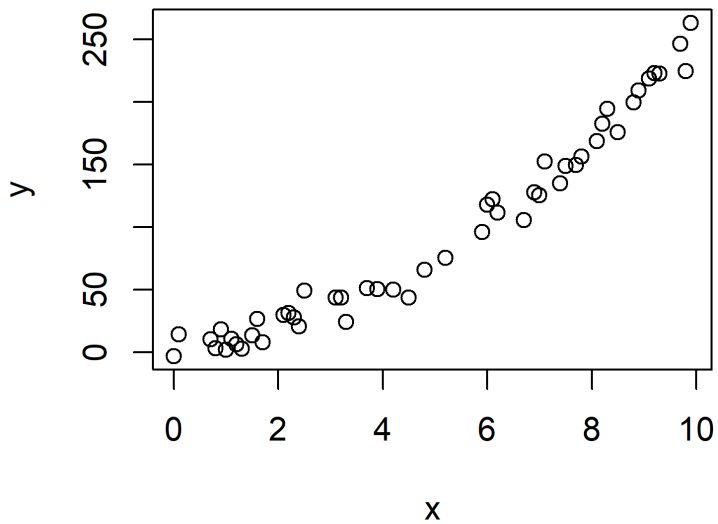


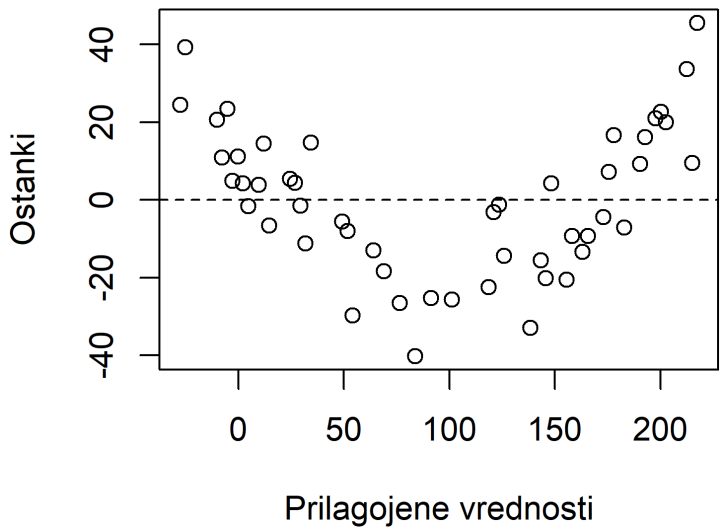


Enostavna linearna regresija

Diagnostika linearnega modela, primer

Nelinearnost

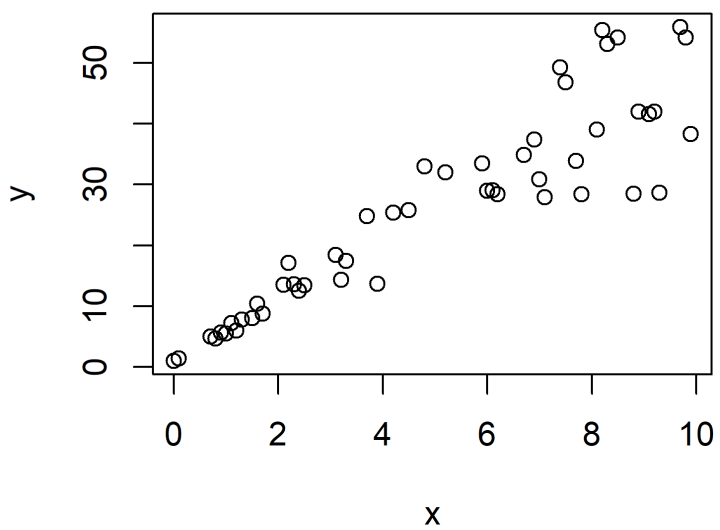


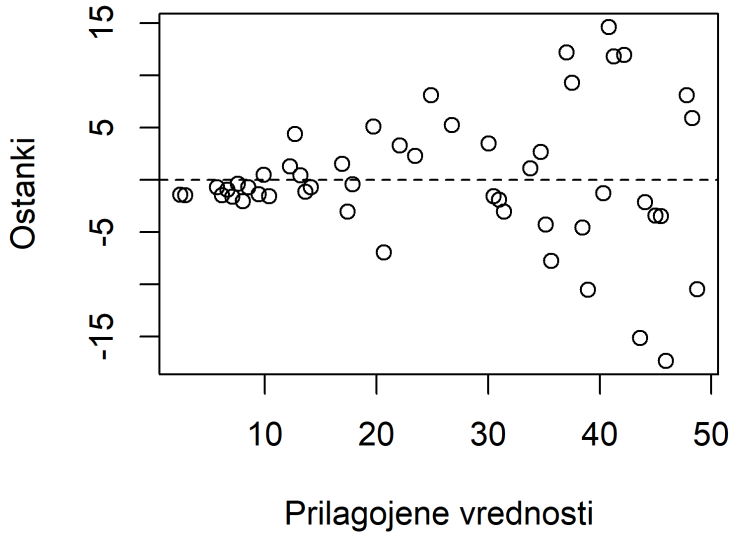


Enostavna linearna regresija

Diagnostika linearnega modela, primer

Nekonstantna varianca, heteroskedastičnost





Enostavna linearna regresija

Porazdelitev cenilk b_0, b_1 in s^2

Ob predpostavki $\varepsilon_i \sim iid\ N(0, \sigma^2)$ in dejstvu, da sta b_0 in b_1 funkciji normalno porazdeljenih spremenljivk, velja, da je tudi njuna porazdelitev aproksimativno (če je n velik) normalna:

$$b_0 \sim N\left(\beta_0,\ \ \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right)\right),$$

$$b_1 \sim N\left(\beta_1,\ \ \frac{\sigma^2}{SS_{xx}}\right).$$

Cenilki varianc parametrov modela $s^2_{b_0}$ in $s^2_{b_1}$ izračunamo tako, da σ^2 zamenjamo z s^2.

Za porazdelitev cenilke variance napak s^2 velja

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$$

Izrek 1.4: Če sta $X \sim N(0,1)$ in $Y \sim \chi^2_n$ neodvisni slučajni spremenljivki, potem velja

$$\frac{X}{\sqrt{Y/n}} \sim t_n$$

Izrek 1.5: Če uporabimo predstavljene lastnosti cenilk in izrek 1.4, lahko pod predpostavko normalne porazdelitve napak ε_i, $i = 1, ..., n$ izpeljemo

$$\frac{b_0 - \beta_0}{s\sqrt{(\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}})}} \sim t_{n-2} \qquad \frac{b_1 - \beta_1}{\frac{s}{\sqrt{SS_{xx}}}} \sim t_{n-2}.$$

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2} \quad \text{in} \quad \frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}.$$

Interval zaupanaja za β_0

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2}$$

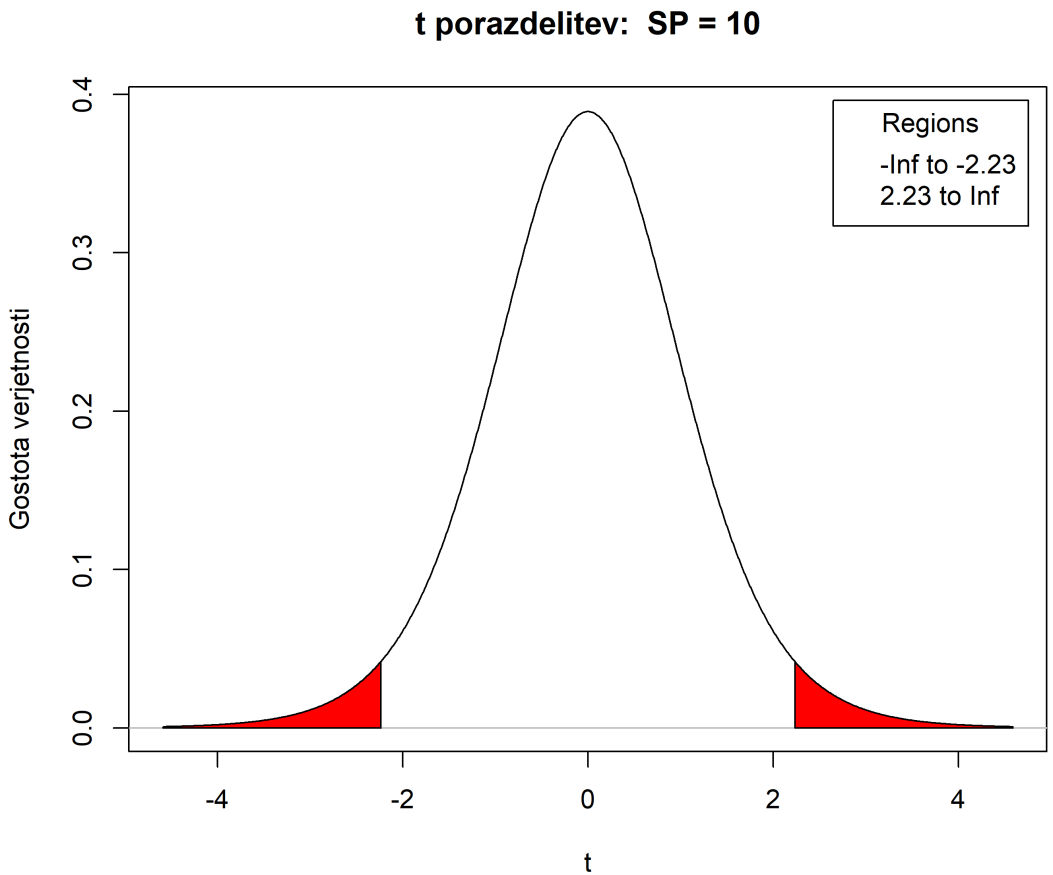
$$P(-|t_{\frac{\alpha}{2};n-2}| \leq \frac{b_0 - \beta_0}{s_{b_0}} \leq |t_{\frac{\alpha}{2};n-2}|) = 1 - \alpha$$

$$P(b_0 - |t_{\frac{\alpha}{2};n-2}|s_{b_0} \leq \beta_0 \leq b_0 + |t_{\frac{\alpha}{2};n-2}|s_{b_0}) = 1 - \alpha$$

$|t_{\frac{\alpha}{2};n-2}|$ je absolutna vrednost $(\alpha/2)$-tega kvantila t-porazdelitve s $SP = n - 2$.

Enostavna linearna regresija

Studentova t-porazdelitev



$t_{0,025;10} = -2,23$ je 0,025-ti kvantil t-porazdelitve s $SP = 10$.

Interval zaupanaja za β_1

$$\frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}$$

$$P(-|t_{\frac{\alpha}{2};n-2}| \leq \frac{b_1 - \beta_1}{s_{b_1}} \leq |t_{\frac{\alpha}{2};n-2}|) = 1 - \alpha$$

$$P(b_1 - |t_{\frac{\alpha}{2};n-2}|s_{b_1} \leq \beta_1 \leq b_1 + |t_{\frac{\alpha}{2};n-2}|s_{b_1}) = 1 - \alpha$$

$|t_{\frac{\alpha}{2};n-2}|$ je absolutna vrednost $(\alpha/2)$-tega kvantila t-porazdelitve s $SP = n - 2$.

Enostavna linearna regresija

Intervali zaupanja za β_0 in β_1

$100(1-\alpha)$ % intervala zaupanja za β_0 in β_1:

$$\left(b_0 - |t_{\frac{\alpha}{2};n-2}|s_{b_0}, \quad b_0 + |t_{\frac{\alpha}{2};n-2}|s_{b_0}\right)$$

$$\left(b_1 - |t_{\frac{\alpha}{2};n-2}|s_{b_1}, \quad b_1 + |t_{\frac{\alpha}{2};n-2}|s_{b_1}\right)$$

Enostavna linearna regresija

Testiranje domnev za β_1

Testiramo ničelno domnevo za β_1

$$H_0 : \beta_1 = \beta \qquad H_1 : \beta_1 \neq \beta.$$

Testna statistika je

$$T = \frac{b_1 - \beta}{\frac{s}{\sqrt{SS_{xx}}}} \sim t_{n-2},$$

Ničelno domnevo zavrnemo pri stopnji značilnosti α, če je

$$T < -|t_{\alpha/2;n-2}| \quad \text{ali} \quad T > |t_{\alpha/2;n-2}|$$

Ob tem je verjetnost, da T leži v območju zavrnitve ničelne domneve, ko je ta pravilna, največ α.

Enostavna linearna regresija

Testiranje domnev za β_1, p-vrednost

Za dani vzorec podatkov izračunamo vrednost testne statistike t, za katero pod ničelno domnevo izračunamo p-**vrednost**

$$p = P(|T| \geq |t| \; | \beta_1 = \beta)$$

Če je $p < \alpha$ ničelno domnevo zavrnemo in če je $p \geq \alpha$ ničelne domenve ne moremo zavrniti.

Enako kot ničelno domnevo z dvostransko alternativno domnevo, lahko testiramo tudi ničelno domnevo z **enostransko alternativno domnevo** $H_1 : \beta_1 < \beta$ ali $H_1 : \beta_1 > \beta$. V tem primeru je p-vrednost polovica p-vrednosti pri testiranju dvostranske alternativne domneve.

Enostavna linearna regresija

Testiranje domnev za β_0

Podobno lahko testiramo ničelno domnevo za β_0

$$H_0 : \beta_0 = \beta \qquad H_1 : \beta_0 \neq \beta$$

Testna statistika je

$$T = \frac{b_0 - \beta}{s\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}})}} \sim t_{n-2}$$

Ničelno domnevo zavrnemo pri stopnji značilnosti α, če je

$$T < -|t_{\alpha/2;n-2}| \quad \text{ali} \quad T > |t_{\alpha/2;n-2}|$$

Ob tem je verjetnost, da T leži v območju zavrnitve ničelne domneve, ko je ta pravilna, največ α.

Enostavna linearna regresija

Testiranje domnev za β_0 in β_1, povzetek modela v R

Ali je zveza med y in x pomembna?

Kakšna je zveza (naraščajoča/padajoča, tesna/šibka)?

Če y ni odvisen od x, je najboljša napoved za y, $\hat{y} = \overline{y}$ in $\beta_1 = 0$.

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

Osnovni povzetek regresijskega modela v R vsebuje rezultat testiranja zgornje domneve in tudi rezultat testiranja ničelne domneve:

$$H_0 : \beta_0 = 0 \qquad H_1 : \beta_0 \neq 0$$

Ta ničelna domneva je redko vsebinsko zanimiva.

Enostavna linearna regresija

Analize variance za regresijski model

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$= SS_{model} + SS_{residual}$$

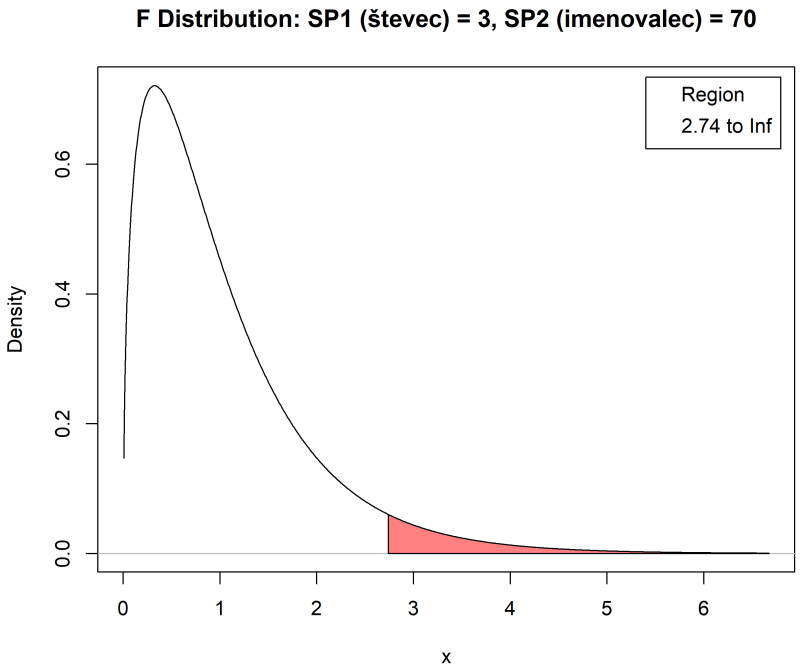
Tabela: Shema tabele ANOVA za enostavni linearni regresijski model

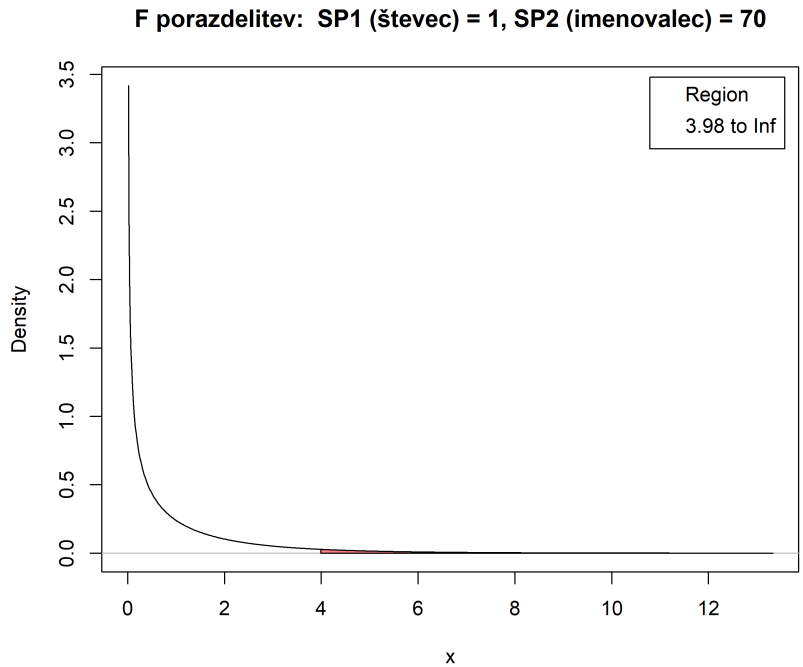
Vir variabilnosti	df	SS	$MS = SS/df$	F
Model	1	SS_{model}	MS_{model}	$MS_{model}/MS_{residual}$
Ostanek (*Residual*)	$n-2$	$SS_{residual}$	$MS_{residual}$	
Skupaj	$n-1$	SS_{yy}		

Enostavna linearna regresija

Analiza variance, F-porazdelitev

Ob predpostavki $\varepsilon_i \sim iid\ N(0, \sigma^2)$ za F-statistiko velja, da je njena ničelna porazdelitev F-porazdelitev s stopinjami prostosti $SP_{model} = k$ in $SP_{residual} = n - 2$.





Iz tabele ANOVA dobimo:

- cenilko za varianco σ^2, ki jo označimo $s^2 = MS_{residual}$. Količino s imenujemo **standardna napaka regresije** (*Residual standard error*).
- F-statistika testira domnevo o ničelnem vplivu napovedne spremenljivke:
  $H_0 : \beta_1 = 0$,
  $H_1 : \beta_1 \neq 0$.

F-test dobi večji pomen v primeru, ko imamo k napovednih spremenljivk v modelu, ker testira ničelno domnevo o hkratni ničnosti vseh parametrov v modelu

$$H_0 : \beta_1 = ... = \beta_k = 0$$

nasproti alternativni domnevi

$$H_1 : \text{vsaj en} \quad \beta_i \neq 0, \quad i = 1, ..., k$$

Izrek 1.6: Če je slučajna spremenljivka X porazdeljena po t-porazdelitvi s stopinjami prostosti v, $X \sim t_v$, potem je slučajna spremenljivka X^2 porazdeljena po F-porazdelitvi s stopinjami prostosti 1 in v, $X^2 \sim F_{1,v}$.

- za testiranje domneve $\beta_1 = 0$ velja

$$T = \frac{b_1}{\frac{s}{\sqrt{SS_{xx}}}} \sim t_{n-2}$$

- če zgornji izraz kvadriramo, dobimo F-statistiko

$$F = \frac{b_1^2 SS_{xx}}{s^2} \sim F_{1,n-2}$$

Enostavna linearna regresija

Povezava med T in F statistiko

- če upoštevamo, $SS_{model} = b_1^2 S_{xx}$ in $s^2 = SS_{residual}/(n-2)$

$$F = \frac{b_1^2 SS_{xx}}{s^2} = \frac{SS_{model}/1}{SS_{residual}/(n-2)} \sim F_{1,n-2}.$$

F-statistika je skalirano razmerje vsote kvadratov odklonov modela in ostanka. Če je SS_{model} veliko večja od $SS_{residual}$, bo F-statistika velika in ničelno domnevo, ki pravi, da regresorji niso uporabni pri napovedovanju odzivne spremenljivke, bomo zavrnili.

Na podlagi ocenjenih parametrov modela lahko izračunamo **povprečno napoved** $\hat{y}(x_0)$, x_0 je izbrana vrednost napovedne spremenljivke.

$$\hat{y}(x_0) = b_0 + b_1 x_0,$$

ob tem je prava napoved

$$\mathbb{E}(y(x_0)) = \beta_0 + \beta_1 x_0$$

Pokažemo lahko, da je porazdelitev povprečne napovedi pri x_0 normalna:

$$\hat{y}(x_0) \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}\right)\right)$$

in velja, da je statistika

$$\frac{\hat{y}(x_0) - \beta_0 - \beta_1 x_0}{s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}} \sim t_{n-2}$$

$100(1-\alpha)\%$ interval zaupanja za povprečno napoved

$$\left(\hat{y}(x_0) - |t_{\frac{\alpha}{2};n-2}|s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}},\quad \hat{y}(x_0) + |t_{\frac{\alpha}{2};n-2}|s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}\right)$$

Z y_0 označimo eno izmed možnih vrednosti odzivne spremenljivke pri x_0

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Napoved za y_0 imenujemo **posamična napoved** in **je enaka povprečni napovedi** $\hat{y}(x_0)$:

$$\mathbb{E}(\hat{y}(x_0) - y_0) = \beta_0 + \beta_1 x_0 - \beta_0 - \beta_1 x_0 = 0$$

Varianca posamične napovedi je:

$$Var(y_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}\right).$$

Enostavna linearna regresija

Napovedovanje, posamična napoved

Podobno kot pri intervalu zaupanja za povprečno napoved lahko definiramo T-statistiko:

$$\frac{\hat{y}(x_0) - y_0}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}} \sim t_{n-2}$$

in $100(1-\alpha)$ % interval zaupanja za posamično napoved je

$$\hat{y}(x_0) \mp |t_{\frac{\alpha}{2};n-2}| s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

- velikost vzorca n
- varianca napak, ocenjena z s^2
- položaj x_0, najmanjša je pri povprečju $\overline{x}$ in narašča s kvadratom razdalje od povprečja.

Enostavna linearna regresija

Primer: intervali zaupanja za napovedi