Diagnostika linearnega modela

Preverjanje predpostavk linearnega modela:

- **linearnost**; enostvana regresija: razsevni grafikon y glede na x; multipla regresija: "grafikoni parcialnih ostankov";
- **pričakovana vrednost napak je 0**, gladilnik na sliki ostankov glede na prilagojene vrednosti se mora čim bolje prilegati abscisi;
- **varianca napak je konstantna** (slika ostankov ali transformiranih standardiziranih ostankov glede na napovedane vrednosti);
- **porazdelitev napak je normalna** (kvantilni graf za standardizirane ostanke);
- **napake so medsebojno neodvisne** (težko preveriti, ustrezni način pridobivanja podatkov, princip slučajnosti; če so podatki izmerjeni v času, ostanke narišemo glede na čas meritve).

Diagnostika linearnega modela

V postopku diagnostike modela uporabljamo grafične prikaze:

- ostankov $\mathbf{e}$
- standardiziranih ostankov $\mathbf{e}_s$
- studentiziranih ostankov $\mathbf{e}_t$
- posebnih točk (regresiski osamelci, vplivne točke, vzvodne točke)

Ostanki

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

Pričakujemo, da imajo ostanki podobne lastnosti kot napake $\varepsilon \sim \mathit{iid}\ N(0, \sigma^2)$: neodvisnost, konstantna varianca, normalna porazdelitev.

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{H}\mathbf{y},$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Ker velja, da so y_i normalno porazdeljene slučajne spremenljivke, to velja tudi za ostanke.

Poiščimo zvezo med ostanki in napakami:

$$\begin{aligned}\mathbf{e} &= (\mathbf{I} - \mathbf{H})\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{H}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{H}\boldsymbol{\varepsilon}.\end{aligned}$$

Ker je $\mathbf{H}\mathbf{X} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{X}$, sledi zveza med ostanki in napakami:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}.$$

Posamezen ostanek e_i zapišemo

$$e_i = (1 - h_{ii})\varepsilon_i - \sum_{j \neq i} h_{ij}\varepsilon_j.$$

$$h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i.$$

Enostavna linearna regresija:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}.$$

Za **vzvode** velja:

- h_{ii}, $i = 1, ...n$ so diagonalni elementi matrike $\mathbf{H}$
- vzvod je odvisen od n, od položaja točke v regresorskem prostoru in od SS_{XX}
- h_{ii} zavzemajo vrednosti med $1/n$ in 1
- $\sum_i^n h_{ii} = k + 1$, kjer je $k + 1$ število parametrov v modelu.

Z večanjem n se elementi matrike $\mathbf{H}$ približujejo vrednosti 0 in ostanki postanejo dobra aproksimacija za napake.

Varianca ostankov $\mathbf{Var}(\mathbf{e})$ je ob predpostavki $Var(\epsilon) = \sigma^2\mathbf{I}$:

$$\mathbf{Var}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\ (\sigma^2\mathbf{I})\ (\mathbf{I} - \mathbf{H})^{\mathrm{T}} = \sigma^2(\mathbf{I} - \mathbf{H})^2 = \sigma^2(\mathbf{I} - \mathbf{H})$$

Varianca ostankov **ni konstantna**, odvisna je od matrike $\mathbf{H}$.

Lastnosti variance ostankov:

- varianca ostankov ni konstantna, odvisna je od matrike **H**, kar pomeni, da je varianca posameznega ostanka je odvisna od položaja točke v regresorskem prostoru;
- kovarianca ostankov $Cov(e_i, e_j) = -\sigma^2 h_{ij}$, za $i \neq j$ je odvisna od vrednosti h_{ij}, njena vrednost se bliža vrednosti 0, ko velikost vzorca n narašča;
- za ostanke e_i velja, da so v absolutnem smislu manjši kot napake ε_i, saj so vzvodi h_{ii} po definiciji pozitivne vrednosti; tudi varianca ostankov $Var(e_i) = \sigma^2(1 - h_{ii})$ je vedno manjša kot varianca napak $Var(\varepsilon_i)$.

- točka z velikim vzvodom ima ostanek z majhno varianco in potencialno lahko predstavlja vplivno točko, ki potegne prilegano premico ali ravnino k sebi, da s tem zagotovi manjšo vrednost ostanka;
- zaradi naštetih lastnosti ostanki niso najboljše vrednosti za diagnostiko modela (standardizirani ostanki, studentizirani ostanki);
- ker velja $Cov(\mathbf{e}, \hat{\mathbf{y}}) = 0$, lahko na podlagi grafa ostankov glede na prilagojene vrednosti preverjamo predpostavko linearnosti zveze med $\mathbf{y}$ in regresorji.

Izrek 2.2: če je varianca napak $Var(\varepsilon) = \sigma^2\mathbf{I}$ so ostanki $\mathbf{e}$ nekorelirani s prilagojenimi vrednostmi odzivne spremenljivke $\hat{\mathbf{y}} = \mathbf{Hy}$ oziroma $Cov(\mathbf{e}, \hat{\mathbf{y}}) = 0$.
Dokaz:

$$\begin{aligned} Cov(\mathbf{e}, \hat{\mathbf{y}}) &= Cov((\mathbf{I} - \mathbf{H})\mathbf{y}, \mathbf{Hy}) \\ &= (\mathbf{I} - \mathbf{H})\ (\sigma^2\mathbf{I})\ \mathbf{H}^{\mathrm{T}} \\ &= \sigma^2(\mathbf{H}^{\mathrm{T}} - \mathbf{H}\mathbf{H}^{\mathrm{T}}) \\ &= 0 \end{aligned}$$

Posledica: razsevni grafikon ostankov glede na prilagojene vrednosti je dobro diagnostično orodje za regresijski model. Če na grafu vidimo odvisnost ostankov od prilagojenih vrednosti, model ne ustreza predpostavkam linearnega modela.

Ker varianca ostankov ni konstantna, je smiselno izračunati **standardizirane ostanke**:

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

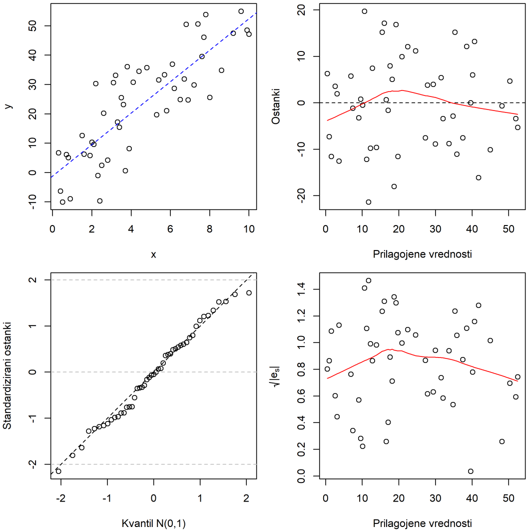
Ker σ v splošnem ne poznamo, jo ocenimo z $\hat{\sigma}$, tako izračunane standardizirane ostanke imenujemo tudi notranje studentizirani ostanki (*internally studentized residuals*).

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

Če so predpostavke modela izponjene, imajo standardizirani ostanki konstantno varianco. Porazdelitev standardiziranih ostankov je približno t_{n-k-1}; če pa je $n >> k$, je porazdelitev približno $N(0, 1)$.

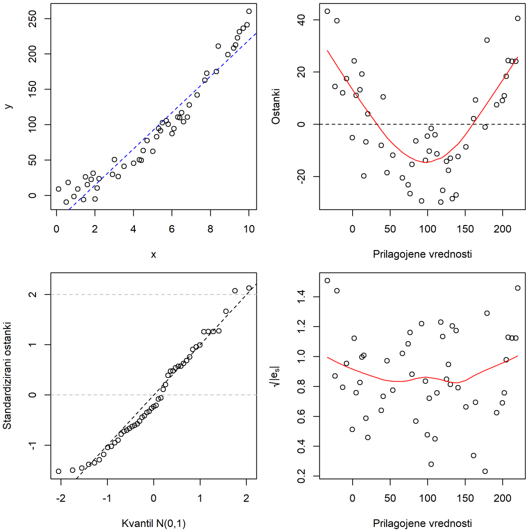
Diagnostika linearnega modela

Primer grafičnih prikazov, ko so predpostavke linearnega modela izpolnjene



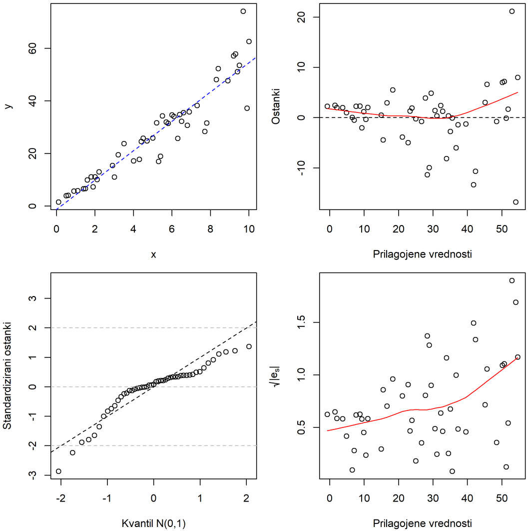
Diagnostika linearnega modela

Predpostavka o linearni zvezi ni izpolnjena



Diagnostika linearnega modela

Predpostavka o konstantni varianci ni izpolnjena



Studentizirani ostanki
Povezanosti med števcem in imenovalcem pri e_{s_i}, $i = 1, ..., n$, se znebimo z izračunom studentiziranih ostankov. Cenilka za σ se izračunana brez upoštevanja i-te točke:

$$e_{t_i} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(-i)} \cdot \sqrt{1 - h_{ii}}} \sim t_{n-k-2}$$

$\hat{\sigma}_{(-i)}$ se izračunana tako, da je v regresijskem modelu i-ta točka izpuščena. Posledično sta števec in imenovalec neodvisna.

Studentizirani ostanki so primernejši za za odkrivanje regresijskih osamelcev kot standardizirani ostanki, saj je $\hat{\sigma}_{(-i)}$ v primeru zelo odstopajoče vrednosti znatno manjša od $\hat{\sigma}$, kar poveča vrednost studentiziranega ostanka.

Graf dodane spremenljivke je prikaz vpliva posameznega regresorja na odzivno spremenljivko ob upoštevanju ostalih regresorjev v modelu.
Za model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, ..., n,$$

graf dodane spremenljivke x_j naredimo na podlagi **ostankov dveh modelov**:
Prvi model za y_i brez regresorja x_j:

$$y_i^{(-j)} = \beta_0^{(j)} + \beta_1^{(j)} x_{i1} + \cdots + \beta_{j-1}^{(j)} x_{i,j-1} + \beta_{j+1}^{(j)} x_{i,j+1} + \beta_k^{(j)} x_{ik} + \varepsilon_i$$

$$e_{i,y}^{(-j)} = y_i - \hat{y}_i^{(-j)}, \quad i = 1, ...n$$

Drugi model za x_j v odvisnosti od ostalih regresorjev:

$$x_{ij}^{(-j)} = \gamma_0 + \gamma_1 x_{i1} + \cdots + \gamma_{j-1} x_{i,j-1} + \gamma_{j+1} x_{i,j+1} + \gamma_k x_{ik} + \varepsilon_i$$

$$e_{i,x_j}^{(-j)} = x_{ij} - \hat{x}_{ij}^{(-j)}, \quad i = 1, ...n$$

Ostanki $e_{i,y}^{(-j)}$ in $e_{i,x_j}^{(-j)}$ predstavljajo vrednosti y in x_j "očiščene" za vpliv ostalih spremenljivk v modelu.

Graf dodane spremenljivke narišemo kot razsevni grafikon za odvisnost $e_{i,y}^{(-j)}$ od $e_{i,x_j}^{(-j)}$ (funkcija `avPlot` iz paketa `car`).

Za premico, ki opisuje odvisnost ostankov $e_{i,y}^{(-j)}$ od $e_{i,x_j}^{(-j)}$ velja:

- naklon premice, je enak oceni parametra b_j iz polnega modela;
- ostanki te premice so enaki ostankom polnega modela;
- standardna napaka naklona te premice je skoraj enaka standardni napaki ocene parametra b_j v polnem modelu (razlikuje se zaradi stopinj prostosti ostanka pri izračunu ocene s^2).

Opisane lastnosti grafa dodane spremenljivke omogočajo diagnostiko linearnega modela z več napovednimi spremenljivkami tudi v kontekstu analize nekonstantne variance in vplivnih točk.

Graf parcialnih ostankov (*Partial Residual Plots*)

Ta grafikon omogoča preverjanje linearnosti oziroma prisotnost nelinearnosti v modelu z več napovednimi spremenljivkami (funkcija `crPlots()` iz paketa `car`).

Za model $y_i = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon_i$ izračunamo **parcialne ostanke** za vsako od napovednih spremenljivk x_j, $j = 1, ...k$:

$$e_{i,x_j} = e_i + b_j x_{ij}.$$

Graf parcialnih ostankov prikazuje parcialne ostanke e_{i,x_j} v odvisnosti od x_{ij}.
Na grafu je tudi gladilnik dobljen z neparametrično regresijo, ki jo izračuna funkcija `lowess()`.
Ta graf pokaže morebitno nelinearnost v zvezi y in x_j, ki je nismo zaobjeli v linearnem modelu.

Če je v model vključena interakcija napovednih spremenljivk, funkcija `crPlots()` ni uporabna. Diagnostiko modela naredimo na podlagi grafov parcialnih ostankov s pomočjo funkcije `Effect()` iz paketa `effects`.

Primer PACIENTI v gradivu, skriptna datoteka `primerP3.R`.

Posebne točke v regresijski analizi so enote, ki zelo odstopajo od ostalih glede na določene kriterije. Te točke prispevajo pomembno informacijo o regresijskem modelu, zato je vedno potrebna njihova analiza.

Pogledali bomo tri vrste posebnih točk:

- **regresijski osamelci**
- **vzvodne točke**
- **vplivne točke**

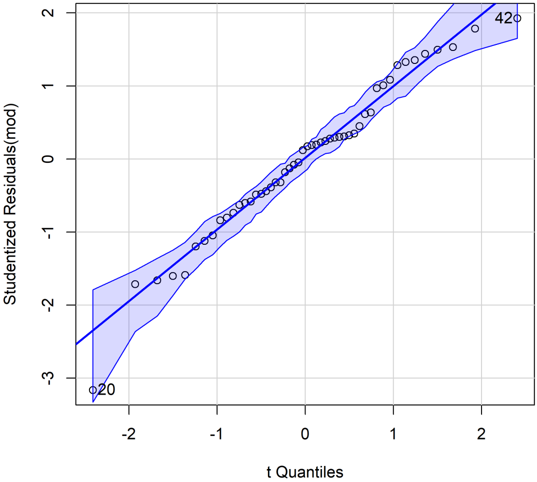
Regresijski osamelec je točka, pri kateri vrednost spremenljivke y_i močno odstopa od pripadajoče napovedane vrednosti $\hat{y}_i$.

Regresijske osamelce ugotavljamo na osnovi studentiziranih ostankov: grafični način ali z modelom.

Funkcija `qqPlot` iz paketa `car` nariše studentizirane ostanke glede na kvantile t_{n-k-2} in pripadajočo 95 % ovojnico, ki je izračunana s parametričnim bootstrap pristopom (Aitkinson, 1985). Točke, ki ležijo izrazito izven ovojnice, so regresijski osamelci.

Diagnostika linearnega modela

Regresijski osamelec, `qqPlot()`



Model za ugotavljanje regresijskih osamelcev (*Mean-shift outlier model*) za i-to točko:

$$y_i = \alpha + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \gamma d_i + \varepsilon_i, \qquad i = 1, ..., n,$$

d_i je umetna spremenljivka z vrednostjo 1 za i-točko in 0 za ostale točke.

Za vsako točko posebej, $i = 1, ..., n$, preverjamo ničelno domnevo, da :

$$H_{0i} : \gamma = 0 \quad i\text{-ta točka ni regresijski osamelec}$$

$$H_{1i} : \gamma \neq 0 \quad i\text{-ta točka je regresijski osamelec}$$

Če velja $\gamma \neq 0$, se presečišče premakne iz α na $\alpha + \gamma$, ob upoštevanju enake odvisnosti y od $(x_1, ..., x_k)$ kot velja za ostale točke.

Teorija pokaže, da je testna statistika pod ničelno domnevo studentizirani ostanek za i-to točko

$$e_{t_i} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{(-i)} \cdot \sqrt{1 - h_{ii}}} \sim t_{n-k-2}$$

Naredimo torej n odvisnih testov zato je treba dobljene p-vrednosti prilagoditi. Funkcija `outlierTest` iz paketa `car` vrne po Bonferroniju popravljene p-vrednosti.

Vzvodne točke so daleč od centra regresorskega prostora, imajo **velik vzvod**, vrednost h_{ii}.

Velja: $\sum_i^n h_{ii} = k+1$, kjer je $k+1$ število parametrov v modelu. Povprečni vzvod je:

$$\overline{h} = \frac{k+1}{n}.$$

Za i-to točko, ki ima vzvod h_{ii} večji od dvakratnika povprečnega vzvoda, pravimo, da je **vzvodna točka**:

$$h_{ii} > 2\overline{h} = 2 \cdot \frac{k+1}{n}.$$

Glede določitve, kako velik mora biti vzvod, da je točka vzvodna, obstoja tudi bolj ohlapno pravilo: $h_{ii} > 3\overline{h}$.

Točka $(y_i, x_{i1}, ...x_{ik})$ **je vplivna**, če se ocene parametrov modela $\mathbf{b}$ ali pa z modelom prilagojene vrednosti $\hat{y}_i$, $i = 1, ..., n$ bistveno spremenijo, če jo izločimo iz modela. Vplivna točka lahko vpliva na statistično sklepanje za parametre modela.

Vplivnost posamezne točke vrednotimo z različnimi merami, ki temeljijo na:

- razlikah $(\mathbf{b}_{(-i)} - \mathbf{b})$, kjer je $\mathbf{b}_{(-i)}$ vektor ocen parametrov v modelu, kjer $i - to$ točko izločimo (**Cookova razdalja**, **DFBETAS**)
- razlikah napovedi $(\hat{y}_i - \hat{y}_{i(-i)})$, $i = 1, ..., n$, kjer je $\hat{y}_{i(-i)}$ napoved v i-ti točki za model, ki $i - te$ točke pri oceni parametrov ne upošteva (**DFFITS**).

Cook (1977) je definiral **Cookovo razdaljo** D_i tako, da meri vpliv i-te točke na **skupno spremembo ocen parametrov** $(\boldsymbol{b}_{(-i)} - \boldsymbol{b})$.

$$D_i = \frac{(\boldsymbol{b}_{(-i)} - \boldsymbol{b})^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} (\boldsymbol{b}_{(-i)} - \boldsymbol{b})}{(k+1)\hat{\sigma}^2}.$$

$\hat{\sigma}^2$ je ocena za varianco napak.

Ta razdalja je osnovana na podlagi skupnega območja zaupanja za vektor parametrov modela $\boldsymbol{\beta}$. Če je Cookova razdalja večjo od 0,5, vektor $\boldsymbol{b}_{(-i)}$ pade izven 50 % skupnega območja zaupanja za $\boldsymbol{\beta}$, za model za vse podatke.

Točka je vplivna, če ima Cookovo razdaljo večjo od , $D_i > 1$.

Pokažemo lahko, da se D_i izrazi s standardiziranim ostankom in vzvodom:

$$D_i = \frac{e_{si}^2}{k+1} \cdot \frac{h_{ii}}{1-h_{ii}}.$$

Točka z veliko vrednostjo standardiziranega ostaneka in hkrati z velikim vzvodom ima velik vpliv na ocene parametrov in posledično tudi na modelske napovedi.

Cookovo razdaljo lahko zapišemo tudi na osnovi prilagojenih vrednosti:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(-i)})^2}{(k+1)\hat{\sigma}^2}.$$

Cookova razdalja je torej skalirana Evklidska razdalja med vektorjem napovedi modela narejenega na vseh podatkih in vektorjem napovedi modela na podatkih, kjer je i-ta točka izločena.

Točke z veliko Cookovo razdaljo identificiramo na četrtem diagnostičnem grafikonu za model. Na razsevnem grafikonu standardiziraniih ostankov in vzvodov sta prikazani izoliniji za Cookovo razdaljo z vrednostma 0.5 in 1.

Primer PADAVINE, `primerP3.R`