4.3 PRESS rezidua

- pokud budeme chtít model použít nejen k vysvětlení vztahu mezi proměnnými, ale také pro predikci, hodila by se míra vyjadřující, jak dobře model predikuje
- šlo by použít IS nebo IP, museli bychom předem znát body, ve kterých chceme predikovat
- nejjednodušší přístup, jak měřit prediktivní přesnost modelu, by byl analýza reziduí pro predikce hodnot v nových $\boldsymbol{x}$
  obecně ale nemáme data y v těchto bodech
- jedna možnost je použít data, která máme k dispozici
  postup: vynecháme jedno pozorování, naladíme model bez tohoto pozorování a porovnáme predikovanou a pozorovanou hodnotu pro vynechané pozorování
- předp., že vynecháme i-té pozorování a označme

  $\widehat{\beta}_{(-i)}$ - odhad β v modelu s vynechaným i-tým pozorováním $(M_{(-i)})$

  $\widehat{y}_{(-i)}$ - predikovanou hodnotu modelem $M_{(-i)}$ v bodě $\boldsymbol{x}_i^T$, tzn. $\widehat{y}_{(-i)} = \boldsymbol{x}_i^T \widehat{\beta}_{(-i)}$

  potom

  $$\widehat{e}_{(-i)} = y_i - \widehat{y}_{(-i)}, \quad i = 1, \dots, n,$$

  nazýváme *i*-té PRESS reziduum

Poznámka 4.3

Otázka je, jak počítat $\widehat{e}_{(-i)}$, $i = 1, \dots, n$

- pro n velké, se to zdá náročný problém, pro každé i je třeba naladit nový model
- naštěstí to nebude nutné, ukážeme totiž $\widehat{e}_{(-i)} = \dfrac{\widehat{e}_i}{1 - h_{ii}}$

Označme $\boldsymbol{x}_i^T$ i-tý řádek matice $\boldsymbol{X}$, $\boldsymbol{X}_{(-i)}$ matici $\boldsymbol{X}$ bez i-tého řádku a $h_{ii} = \boldsymbol{H}_{ii}$.

Věta 4.1

Jestliže $h_{ii} \neq 1$, potom
$$\left(\boldsymbol{X}_{(-i)}^T \boldsymbol{X}_{(-i)}\right)^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{1 - h_{ii}}.$$

Důkaz.

Věta z LA: (Sherman-Morrison-Woodbury)

Nechť $\boldsymbol{A}$ je $n \times n$ invertibilní matice a nechť $\boldsymbol{z}$ je $n \times 1$ sloupcový vektor. Jestliže $\boldsymbol{z}^T \boldsymbol{A}^{-1} \boldsymbol{z} \neq 1$, potom matice $\boldsymbol{B} = \boldsymbol{A} - \boldsymbol{z}\boldsymbol{z}^T$ je invertibilní a platí

$$\boldsymbol{B}^{-1} = \boldsymbol{A}^{-1} + \frac{\boldsymbol{A}^{-1} \boldsymbol{z}\boldsymbol{z}^T \boldsymbol{A}^{-1}}{1 - \boldsymbol{z}^T \boldsymbol{A}^{-1} \boldsymbol{z}} \, .$$

Věta 4.2

Nechť $\widehat{e}_{(-i)}$ je i-té PRESS reziduum. Potom $\widehat{e}_{(-i)} = \dfrac{\widehat{e}_i}{1 - h_{ii}}, \quad i = 1, \ldots, n.$

Důkaz.

Věta 4.3

1) Nechť $\widehat{\beta}_{(-i)}$ značí *LSE* parametru β v modelu bez i-tého pozorování. Potom platí

$$\widehat{\beta} - \widehat{\beta}_{(-i)} = \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\widehat{e}_i}{1 - h_{ii}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\widehat{e}_{(-i)}.$$

2) Pro součet residuálních čtverců $SSE_{(-i)}$ v modelu bez i-tého pozorování platí

$$SSE_{(-i)} = \sum_{j=1}^{n} \widehat{e}_j^2 - \frac{\widehat{e}_i^2}{1 - h_{ii}}.$$

Důkaz.

Důsledek 4.1

V modelu $(**)$ s $m+1$ parametry β a bez i-tého pozorování platí

$$\mathrm{E}\left[SSE_{(-i)}\right] = (n-m-2)\sigma^2, \quad \text{to znamená} \quad \widehat{\sigma}^2_{(-i)} = \frac{SSE_{(-i)}}{n-m-2} \quad \text{je nestranný odhad } \sigma^2.$$

Dále pak

$$\widehat{\sigma}^2_{(-i)} = \frac{(1-h_{ii})(n-m-1)s_n^2 - \widehat{e}_i^2}{(1-h_{ii})(n-m-2)} = \frac{1}{n-m-2}\left(SSE - \frac{\widehat{e}_i^2}{1-h_{ii}}\right),$$

kde $s_n^2 = \frac{1}{n-m-1} SSE$ (pro plný model).

Důkaz.

Poznámka 4.4

- dá se ukázat, že $SSE_{(-i)}$ a $\widehat{e}_i$ jsou nezávislé náhodné veličiny
- protože $\dfrac{SSE_{(-i)}}{\sigma^2} \sim \chi^2(n-m-2)$ a $\dfrac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$
- dostaneme $\dfrac{\widehat{e}_i}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} \sim t(n-m-2)$

Tvrzení 4.1

Uvažujme model $(**)$, kde $h(\boldsymbol{X}) = m+1$ a $\boldsymbol{e} \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$. Nechť pro $i \in \hat{n}$ platí, že $h_{ii} \neq 1$. Potom i-té (externě) studentizované reziduum

$$\widehat{t}_i \sim t(n-m-2).$$

Poznámka 4.5

- $\widehat{t}_i$ lze použít pro test hypotézy, zda je i-té pozorování odlehlé (outlier), tedy

$$H_0: \ i\text{-té pozorování není odlehlé v modelu } M \quad \times \quad H_1: \ i\text{-té pozorování je odlehlé v } M,$$

- kde odlehlé značí odlehlé vzhledem k $M: \ \boldsymbol{Y} \sim N_n(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$:
  a) střední hodnota i-tého pozorování se nerovná té dané modelem,
  b) pozorovaná hodnota Y_i je neobvyklá za platnosti M.
- H_0 zamítneme, pokud $\quad |\widehat{t}_i| > t_{1-\frac{\alpha}{2}}(n-m-2)$
- pokud test použijeme na všechna pozorování, je potřeba aplikovat nějakou korekci na vícenásobné testování, např. Bonferroni

Poznámka 4.6 (Vztah mezi $\widehat{e}_{(-i)}$ a $\widehat{t}_i$)

- $\widehat{e}_{(-i)} = \dfrac{\widehat{e}_i}{1-h_{ii}} \quad \Rightarrow \quad \mathsf{E}\,\widehat{e}_{(-i)} = 0, \qquad \operatorname{Var}\widehat{e}_{(-i)} = \dfrac{\sigma^2}{1-h_{ii}}$
- standardizované PRESS reziduum $\dfrac{\widehat{e}_{(-i)}}{\sqrt{\operatorname{Var}\widehat{e}_{(-i)}}} = \dfrac{\frac{\widehat{e}_i}{1-h_{ii}}}{\frac{\sigma}{\sqrt{1-h_{ii}}}} = \dfrac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}} = r_i$
- pokud použijeme $\widehat{\sigma}^2_{(-i)}$ jako odhad σ^2, pak studentizovaná PRESS rezidua

$$\frac{\widehat{e}_i}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} = \widehat{t}_i$$

Poznámka 4.7

- $\widehat{e}_{(-i)} = \dfrac{\widehat{e}_i}{1 - h_{ii}} \Rightarrow$ pokud i-té pozorování má velké h_{ii}, bude $\widehat{e}_{(-i)}$ mnohem větší, než $\widehat{e}_i$,

  - pozorování s velkým h_{ii} jsou dobře modelována, ale měřeno $\widehat{e}_{(-i)}$ mohou špatně predikovat

  - to je další ukázka fit/prediction dilema

- stejný efekt nastává také pro $\quad \widehat{\beta} - \widehat{\beta}_{(-i)} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i \widehat{e}_{(-i)}$

  rozdíl může být „malý“, pokud je „fit“ dobrý, ale může být také „velký“, pokud je h_{ii} velké.

Míry influence

- i pro perfektní model mohou dva různé vzorky $(\mathbf{x}, \mathbf{y})$ a $(\mathbf{x}', \mathbf{y}')$ vést k různým závěrům
- většinou máme k dispozici jen originální data
- bude nás zajímat vliv i-tého řádku $\boldsymbol{X}$ na model
- už víme, že velké h_{ii} indikuje, že i-té pozorování má velký vliv a velká rezidua naznačují možnou neadekvátnost modelu
- míry, které zavedeme, budou kombinovat tyto dva faktory
- přístup z PRESS reziduí, tzn. jak velký vliv má vynechání i-tého pozorování na $\widehat{\beta}$ a $\widehat{\boldsymbol{y}}$

DFBETAS:

vliv vynechání i-tého pozorování na odhad $\widehat{\beta}$ měří rozdíl

$$\widehat{\beta} - \widehat{\beta}_{(-i)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \frac{\widehat{e}_i}{1 - h_{ii}},$$

bude tedy základem pro naši analýzu

a) vliv i-tého pozorování na $\widehat{\beta}_j$:

- $\widehat{\beta}_j - \widehat{\beta}_{(-i)j} = \dfrac{r_{ji}\,\widehat{e}_i}{1-h_{ii}}$, kde r_{ji} je (j,i)-tý prvek matice $\boldsymbol{R} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
- i-té pozorování budeme považovat za influenční na β_j, pokud $\widehat{\beta}_j - \widehat{\beta}_{(-i)j}$ bude velké
- protože $\widehat{\beta}_j$ je náhodná veličina, „velké" bychom měli měřit relativně vzhledem k $s.d.(\widehat{\beta}_j) = \sigma\sqrt{v_j}$, kde $v_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}$
- pokud ji odhadneme pomocí $\widehat{\sigma}_{(-i)}\sqrt{v_j}$, dostaneme definici

$$\mathrm{DFBETAS}_{j,i} = \frac{\widehat{\beta}_j - \widehat{\beta}_{(-i)j}}{\widehat{\sigma}_{(-i)}\sqrt{v_j}} = \frac{r_{ji}\widehat{e}_i}{\sqrt{v_j}\widehat{\sigma}_{(-i)}(1-h_{ii})} = \frac{r_{ji}}{\sqrt{v_j}}\frac{\widehat{t}_i}{\sqrt{1-h_{ii}}},$$

  kde $\widehat{t}_i$ je (ext.) studentizované reziduum
- kombinuje efekt velkého rezidua $\widehat{t}_i$ a velkého h_{ii}
- jedna možnost pro limitní hodnoty: i-té pozorování je považováno za influenční na odhad β_j, pokud

$$|\mathrm{DFBETAS}_{j,i}| > \frac{2}{\sqrt{n}}$$

- máme $(m+1)\times n$ hodnot pro srovnání, zjednodušíme

b) vliv i-tého pozorování na celý vektor $\widehat{\beta}$:

- použití nějaké normy na vektor $\widehat{\beta} - \widehat{\beta}_{(-i)}$
- Cook navrhl
$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_{(-i)})^T \boldsymbol{M} (\widehat{\beta} - \widehat{\beta}_{(-i)})}{(m+1)c},$$
  kde $\boldsymbol{M}$ je PD matice a c normalizační konstanta
- nejužívanější volba je $\boldsymbol{M} = \boldsymbol{X}^T \boldsymbol{X}$ a $c = s_n^2$
- Cookova vzdálenost:
$$D_i = \frac{(\widehat{\beta} - \widehat{\beta}_{(-i)})^T \boldsymbol{X}^T \boldsymbol{X} (\widehat{\beta} - \widehat{\beta}_{(-i)})}{(m+1)s_n^2}$$
- dosazením dostaneme
$$D_i = \frac{1}{(m+1)s_n^2}\left(\frac{\widehat{e}_i}{1-h_{ii}}\right)^2 \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i = \frac{1}{m+1} \frac{h_{ii}}{1-h_{ii}} \frac{\widehat{e}_i^2}{s_n^2(1-h_{ii})}$$
- výpočetní tvar potom je
$$D_i = \frac{\widehat{r}_i^2}{m+1}\left(\frac{h_{ii}}{1-h_{ii}}\right) \qquad (\widehat{r}_i \text{ jsou interně studentizovaná (standardizovaná) rezidua})$$

Poznámka 4.8

- $100(1-\alpha)\%$ simultánní IS pro β je:

$$C(\alpha) = \left\{ \beta \,\middle|\, \frac{(\widehat{\beta}-\beta)^T \boldsymbol{X}^T \boldsymbol{X}(\widehat{\beta}-\beta)}{(m+1)s_n^2} \leq \mathrm{F}_{1-\alpha}(m+1, n-m-1) \right\}$$

- tzn. $\widehat{\beta}_{(-i)} \in C(\alpha) \quad \Leftrightarrow \quad D_i \leq \mathrm{F}_{1-\alpha}(m+1, n-m-1)$
- to je motivace pro **RULE OF THUMB**:

i-té pozorování je influenční, jestliže $D_i > F_{\frac{1}{2}}(m+1, n-m-1)$

(pro většinu m, n je $\mathrm{F}_{\frac{1}{2}} \approx 1$, zjednodušení pravidla $D_i > 1$)

Poznámka 4.9

Také platí, že

$$D_i = \frac{(\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{y}}_{(-i)})^T(\widehat{\boldsymbol{y}} - \widehat{\boldsymbol{y}}_{(-i)})}{(m+1)s_n^2},$$

tzn. D_i se dá chápat jako míra influence na celkovou predikci

DFFITS: vliv i-tého pozorování na $\widehat{y}_i$

$$\mathrm{DFFITS}_i = \frac{\widehat{y}_i - \widehat{y}_{(-i)}}{\widehat{\sigma}_{(-i)}\sqrt{h_{ii}}} = \ldots = \widehat{t}_i\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

RULE OF THUMB: i-té pozorování je influenční, pokud $|\mathrm{DFFITS}_i| > 3\sqrt{\frac{m+1}{n-m-1}}$

Poznámka 4.10 (Míry influence v R)

- DFBETAS - `dfbetas()`, DFFITS - `dffits()`,
- Cookova vzdálenost D_i - `cooks.distance()`, potenciál h_{ii} - `hatvalues()`
- vše shrnuje funkce `influence.measures()`
- používaná pravidla: i-té pozorování je označeno za influenční, pokud

$$|\mathrm{DFBETAS}_{j,i}| > 1, \quad |\mathrm{DFFITS}_i| > 3\sqrt{\frac{m+1}{n-m-1}}, \quad D_i > F_{\frac{1}{2}}(m+1, n-m-1), \quad h_{ii} > 3\frac{m+1}{n}.$$