REGRESNÍ DIAGNOSTIKA

Chemometrie I, © David MILDE

Regresní diagnostika

- Obsahuje postupy k posouzení:
  - kvality dat pro regresní model (přítomnost vlivných bodů),
  - kvality modelu pro daná data,
  - splnění předpokladů pro MNČ (či jinou metodu odhadu).
- Společné posouzení všech 3 výše uvedených bodů nám umožňuje studium tzv. regresního tripletu.

Regresní triplet: data + regresní model + metoda odhadu

- Z praktického hlediska (využití software) budeme regresní diagnostiku dělit na 2 části:
  - metody analýzy vlivných bodů,
  - metody pro odhalení porušení předpokladů MNČ a posuzování vhodnosti modelu.

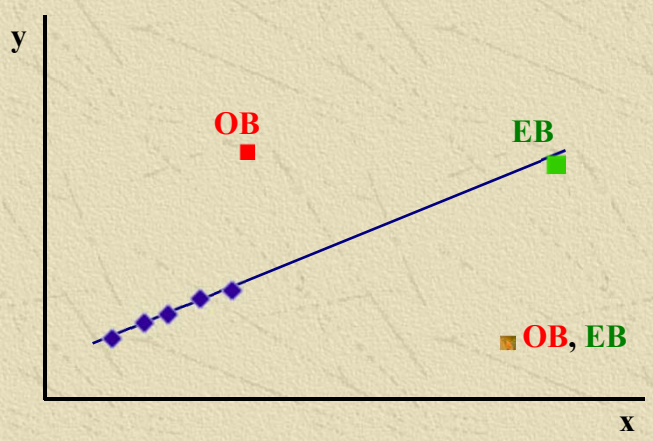
Regresní diagnostika

- Základní rozdíl mezi regresní diagnostikou a „klasickými statistickými testy“ používanými v regresi je v tom, že není třeba přesně formulovat alternativní hypotézu.
- Regresní diagnostika se tak blíží EDA, a umožňuje interaktivní zásah uživatele, který zná „svá data“ lépe než software.
- Tím je omezen vznik formálních regresních modelů, které nemají fyzikální smysl a jsou v praxi obvykle jen omezeně použitelné.

Kvalita dat: vlivné body

- Vlivné body ovlivňují výsledek statistické analýzy tím, že zkreslují regresní model či zvyšují rozptyl.
- Lze je rozdělit do 3 skupin:
  - hrubé chyby – důsledek chyb při manipulaci s daty,
  - body s vysokým vlivem – spolehlivě změřené body rozšiřující predikční schopnost regresního modelu,
  - zdánlivě vlivné body – jeví se jako vlivné, protože byl zvolen nevhodný regresní model.
- Podle místa výskytu se dělí na:
  - odlehlé body (OB) – liší se v hodnotách závisle proměnné,
  - extrémní body (EB) – liší se v hodnotách nezávisle proměnné,
  - kombinace OB a EB, o jejich výsledném vlivu spíše rozhoduje to, že jsou EB.

Kvalita dat: vlivné body



Indikace vlivných bodů: statistická analýza reziduí

- Reziduum je vyčíslená hodnota z regresního modelu a používá se při posuzování kvality modelu i kvality dat.

1. Klasické reziduum $e_i = y_i - y_{i,reg}$
2. Normované reziduum $e_{Ni} = e_i/\sigma$
3. Standardizované reziduum (e_{Si}) – slouží k identifikaci heteroskedasticity
4. Jackknife reziduum (e_{Ji}) – identifikuje OB
5. Predikované reziduum (e_{Pi}) – identifikuje OB

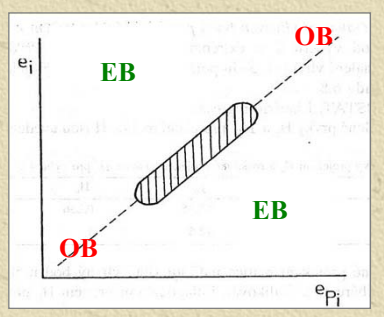
Indikace vlivných bodů: analýza vlivu pomocí indexů

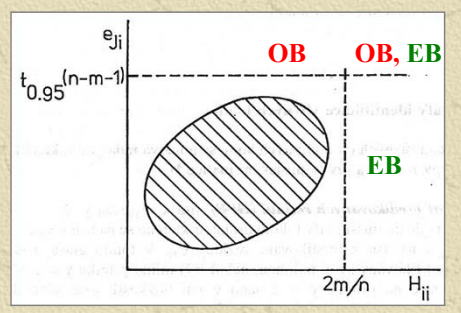
- Velké množství diagnostik vlivných bodů, které jsou založeny na sledování změn, ke kterým dojde při vypuštění jednotlivých bodů a jejich „dopočtení“ z regresního modelu.
- Cookova vzdálenost D_i: je-li $D_i > 1$, bod je vlivný.
- Atkinsonova vzdálenost: modifikace Cookovy vzdálenosti se zvýrazněnou citlivostí na EB.
- Diagonální prvky projekční matice H_{ii}:
  - indikují přítomnost EB, které nezachytí analýza reziduí,
  - $\mathbf{H = X(X^TX)^{-1}X^T}$
- V software se používá barevné zvýraznění bodů identifikovaných jako vlivné.

ACH/CHEX1, 2010

Grafy identifikace vlivných bodů

- Graf predikovaných reziduí (GPR)
- Williamsův graf

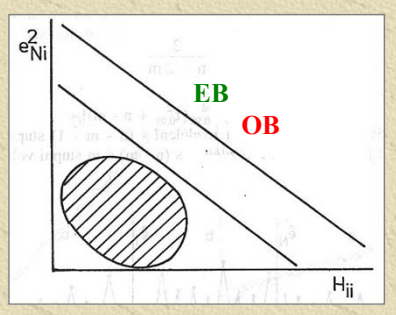


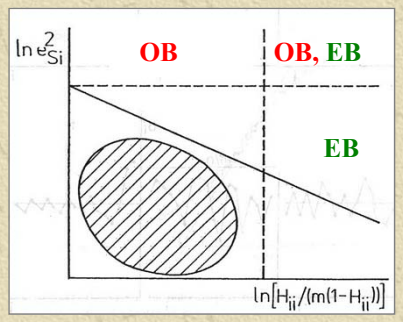


ACH/CHEX1, 2010

Grafy identifikace vlivných bodů

- Pregibonův graf (PG) – nerozlišuje EB od OB
- McCullohův-Meeterův graf (MMG)





Grafy identifikace vlivných bodů

- L-R graf (osa x: H_{ii}, osa y: e_{Ni}^2)
  - Hyperboly znázorňují isolinie stejného vlivu.
  - Podle polohy vůči 3 křivkám lze data rozdělit na slabě vlivná, vlivná a silně vlivná.
- Q-Q graf (osa x: kvantil N(O, 1), osa y: reziduum)
  - Lze konstruovat pro různá rezidua.
  - Kromě vlivných bodů slouží i k posouzení normality reziduí.
- Graf Cookovy vzdálenosti.
- Graf Atkinsonovy vzdálenosti.
- Graf diagonálních prvků projekční matice H.

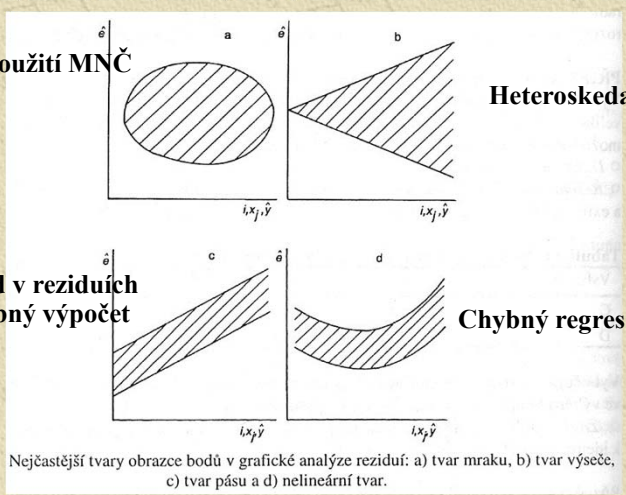
Grafická analýza reziduí

Vhodné použití MNČ

Heteroskedasticita

Trend v reziduích
Chybný výpočet

Chybný regresní model



Nejčastější tvary obrazce bodů v grafické analýze reziduí: a) tvar mraku, b) tvar výseče, c) tvar pásu a d) nelineární tvar.

Ověření předpokladů MNČ
Testování regresního tripletu

- Statistická významnost regresního modelu: F_R test významnosti regrese – testuje, zda použitý model je lepší než „žádný“ model.
  - *viz. přednáška o testování hypotéz v LR*
- Multikolinearita: Scottovo kritérium multikolinearity SC – testuje, zda mezi nezávisle proměnnými není příliš velká kolinearita, která zvyšuje výrazně rozptyl parametrů regresního modelu.
  - *viz. přednáška o polynomické regresi*
- Závislost/trend reziduí: neparametrický test ověřuje přítomnost závislostí, které nejsou postihnuty modelem – posouzení na základě počtu změn +/- reziduí.

Ověření předpokladů MNČ
Testování regresního tripletu

- Heteroskedasticita = nekonstantnost rozptylu: Cook-Weisbergův test; CW se srovnává s $\chi_{krit}^2(1)$.
  - Je-li CW > χ_{krit}^2 – je prokázána heteroskedasticita.

$$CW = \frac{\left[\sum_{i=1}^{n}(y_i-\bar{y})\cdot e_i^2\right]^2}{2\sigma^4\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

- Heteroskedasticitu lze odhalit i v grafu heteroskedasticity (osa x: $(1-H_{ii})y_i$, osa y: e_{Si}^2) $\Rightarrow$ klínový tvar bodů v grafu.
- V přítomnosti heteroskedasticity je třeba uvažovat o použití vah = metodě vážených nejmenších čtverců.

Ověření předpokladů MNČ
Testování regresního tripletu

- Normalita reziduí: Jarque-Bearův test; JB se srovnává s $\chi_{krit}^2(2)$.
  - Je-li JB < χ_{krit}^2 – je prokázána normalita.
  - Test je založen na posouzení statistického rozdělení reziduí.

$$JB = n.\left(\frac{g_1}{6}+\frac{(g_2-3)^2}{24}\right)$$

- Normalitu reziduí lze odhalit i v Q-Q grafech reziduí.

Ověření předpokladů MNČ
Testování regresního tripletu

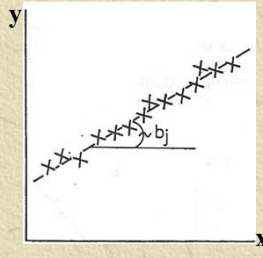
- Autokorelace – v LR bývá důsledkem vynechání významné proměnné související s y: Waldův test test; WA se srovnává s $\chi_{krit}^2(1)$.
  - Je-li WA > χ_{krit}^2 – je prokázána autokorelace.
  - Testuje přítomnost autokorelace chyb na základě reziduí.
- Často se používá i Durbin-Watsonův test, který také ověřuje přítomnost autokorelace na základě reziduí.
  - $0 \leq DW < 2$ a $2 < DW < 4$ potvrzuje autokorelaci.
  - $DW \approx 2$ autokorelace není.

$$WA = \frac{n\rho_1^2}{1-\rho_1^2}$$

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Ověření předpokladů MNČ
Kvalita regresního modelu

- Pomocí rozptylového grafu: y = f(x).
- Pomocí parciálních regresních grafů (zejména u vícerozměrné LR).
  - Závislost y na zvolené x_i s eliminací vlivu ostatních nezávisle proměnných x.
  - Závislost je lineární pouze v případě, že model je správný.



Směrnice přímky = b_i

Ověření předpokladů MNČ Kvalita regresního modelu

- Pomocí charakteristik vhodnosti modelu AIC, MEP, R_p.
- Při porovnávání regresních modelů hledáme MEP a AIC minimální a R_p^2 maximální.
- **Střední kvadratická chyba predikce** - MEP (Mean Error of Prediction)
  - MEP využívá predikce $y_{reg,i}$ z odhadu, při jehož konstrukci byla informace o i-tém bodu vypuštěna. Jde tedy o chybu i-tého bodu závisle proměnné spočítanou regresí právě s vyloučením i-tého bodu.

$$MEP = \frac{1}{n}\sum_{i=1}^{n}\frac{e_i^2}{(1-H_{ii})^2}$$

Ověření předpokladů MNČ Kvalita regresního modelu

- **Predikovaný koeficient determinace** R_p^2 - získáme pokud při výpočtu R^2 použijeme MEP místo RSC, je citlivější na vybočující body než klasický R^2.
  - QC Expert používá predikovaný korelační koeficient R_p.

$$R_p^2 = 1 - \frac{n\,MEP}{\sum_{i=1}^{n} y_i^2 - n\,\bar{y}}$$

- **Akaikovo informační kriterium** AIC – je kritérium kvality regrese vycházející z RSC penalizovaného počtem proměnných n.

$$AIC = n \cdot \ln\left(\frac{RSC}{n}\right) + 2m$$

Výstavba lineárního regresního modelu

1. Návrh modelu (co nejjednodušší předběžný model).
2. Předběžná analýza dat (posouzení R, AIC, MEP, R_p^2, …).
3. Regresní diagnostika zaměřená zejména na kvalitu dat.
4. Konstrukce zpřesněného regresního modelu (případné použití jiných metod odhadu než je MNČ).
5. Posouzení kvality modelu s využitím testů regresního tripletu.
6. Tvorba konečného regresního modelu.