Výběr regresního modelu

- budeme se zabývat výběrem nejvhodnějších nez. regresorů
- špatná specifikace modelu (použití jiného než skutečného modelu) má dva hlavní důsledky:

  1) při vynechání některých proměnných modelu, jsou odhady par. ostatních proměnných vychýlené

  2) pokud je v modelu příliš mnoho proměnných, jsou obecně rozptyly odhadů pro ostatní proměnné velké

  (výběr modelu: trade-off mezi vychýleností a přesností)

- volba „nejlepšího" modelu je hledání kompromisu mezi dvěma kritérii

  a) přesnost modelu   b) jednoduchost modelu (parsimony)

(- ideální model by měl mít nejmenší možný počet regresorů, který umožňuje adekvátní interpretaci (nebo predikci))

- obecně neexistuje jednoznačný nejlepší model ani jednoznačná statistická procedura, jak ho najít

POZN: důsledky vynechání proměnných ze skutečného, a bohužel neznámého, modelu

předp. $Y = X\beta + e$ (true model) a $\beta = (\beta_1^T, \beta_2^T)^T$

$\beta_1 = (\beta_0, \dots, \beta_q)^T, \quad \beta_2 = (\beta_{q+1}, \dots, \beta_m)^T$

podobně lze rozepsat $X = [X_1 \ X_2]$ a model může mít tvar

$$Y = X_1 \beta_1 + X_2 \beta_2 + e \quad \to \quad \hat{\beta} = (X^T X)^{-1} X^T Y$$

redukovaný model $Y = X_1 \beta_1 + e \quad \to \quad \hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y$

– víme, že $\hat{\beta}$ je nestranný odhad β, ale $\hat{\beta}_1$ není nestranný o. β_1

$$E\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T \underbrace{EY}_{= X_1\beta_1 + X_2\beta_2} = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 = \beta_1 + A \beta_2$$

obecně $A \neq 0$, tedy $E\hat{\beta}_1 \neq \beta_1$

– pro rozptyl platí $Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ $\quad Cov(\hat{\beta}_1) = \sigma^2 (X_1^T X_1)^{-1}$

ozn. $\hat{\beta}_1^*$ LSE β_1 v modelu s $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$

dá se ukázat $Cov(\hat{\beta}_1^*) - Cov(\hat{\beta}_1)$ je PSD, tzn. že

rozptyly $\hat{\beta}_1^*$ budou obecně větší než rozptyly $\hat{\beta}_1$

testy vynechání proměnných zvyšují přesnost odhadů $\hat{\beta}_0, \dots, \hat{\beta}_r$

– predikované hodnoty $\hat{Y} = HY$ $\quad \hat{Y}_1 = H_1 Y$

$$E\hat{Y}_1 = \underbrace{H_1}_{= X_1 (X_1^T X_1)^{-1} X_1^T} EY = H_1 (X_1 \beta_1 + X_2 \beta_2) = X_1 \beta_1 + X_1 A \beta_2$$

pro $A \neq 0$ je $\hat{Y}_1$ vychýlený odhad Y

Kritéria pro ~~výběr~~ porovnávání modelů

předpokládejme, že máme k dispozici T proměnných (regresorů) včetně interceptu

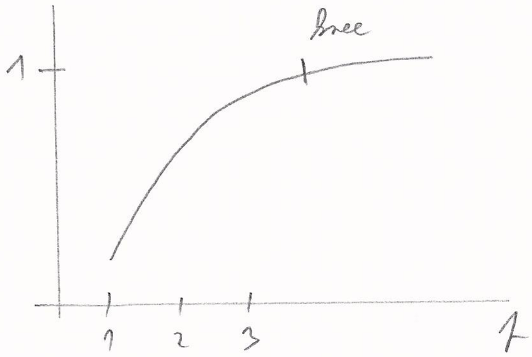
a uvažujeme podmnožinu t proměnných (včetně interceptu)

$(t = m+1)$

A) koeficient vícenásobné determinace R^2

$$R_p^2 = \frac{SSR_p}{SST} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SSE_p}{SST}$$

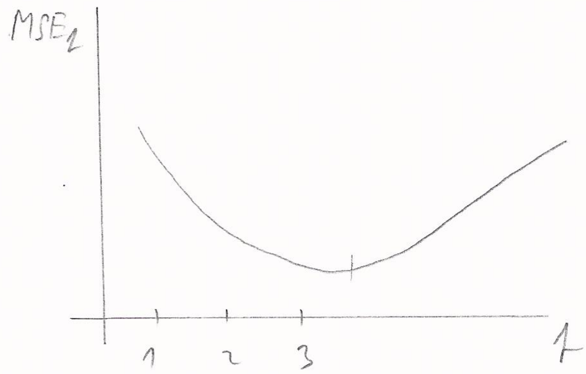
- při použití je třeba si uvědomit, že R_p^2 je rostoucí funkcí počtu regresorů p (SST je konst.), tedy maxima nabyde pro $p = T$
- hledáme tedy model ve kterém přidání dalšího regresoru už nezpůsobí podstatný nárůst R_p^2



- často se používá

$$\bar{R}_p^2 = 1 - \frac{\frac{SSE_p}{n-p}}{\frac{SST}{n-1}}$$

B) (R)MSE



$$MSE_p = \frac{SSE_p}{n-p} = s_n^2$$

$$RMSE_p = s_n$$

C) F - test pro vnořené modely

pro $X = (X_1, X_2)$ $\quad \beta = \binom{\beta_1}{\beta_2}$ $\quad$ umíme otestovat

$H_0 : \beta_2 = 0$ $\quad$ pomocí F testu

extra sum of squares princip

R : anova( ) : pozor na vnořenost modelů
záleží na pořadí v jakém přidáváme regresory do modelu

D) Mallows C_p, AIC, BIC

- kritéria berou více v potaz počet použitých regresorů
- lze je použít i pro nevnořené modely !

penalizace

• Mallows C_p

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + \cancel{2(p+1)}\ 2p \qquad \hat{\sigma}^2 = \frac{SSE_T}{n-T}$$

Vlastnosti C_p: 1) snadno se počítá, SSE_p, $\hat{\sigma}^2$ implementované

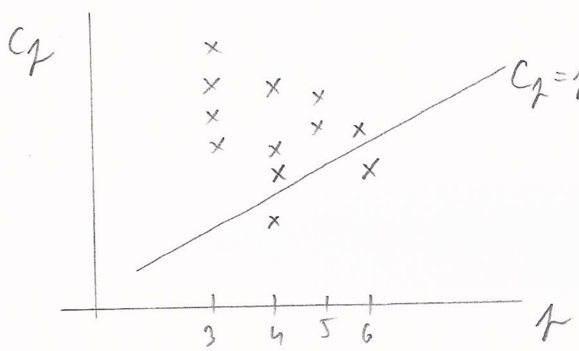
2) pokud je $\hat{\sigma}^2$ konzistentní odhad σ^2 (nezávisející na p)
má C_p následující interpretaci:
porovnává co nelze vysvětlit pomocí modelu s p a T parametry
zohledňuje počet dostupných dat a penalizuje počet parametrů, které je třeba odhadnout

3) při navyšování počtu regresorů
$\hat{\sigma}^2$ je konst., SSE_p klesá, p roste
(C_p se snaží vyvážit dvě protichůdná kritéria)

4) $C_T = T$

5) pokud je správný model s p parametry, dá se ukázat, že
$C_p \approx p$ pro $n \gg T$

6) v praxi se volí model s nejmenším C_p ve skupině modelů splňujících $C_p \approx p$



body nejblíže přímky indikují dobré modely

POZN: nevýhoda - pro dobrou interpretaci je třeba spočítat C_p pro všechny nebo většinu podmnožin regresorů

• Akaikeho informační kritérium AIC

obecná definice je

$$AIC = -2\ell(\hat{\theta}) + 2p^*$$

$\hat{\theta}$ je MLE odhad v modelu
ℓ je log-věrohodnostní funkce
p^* - počet par., které je třeba odhadovat
($p^* = p+1$ počítáme i σ^2)

pro náš model LR

$$L(\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)}$$

$$\ell(\beta,\sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\frac{(y-X\beta)^T(y-X\beta)}{\sigma^2}$$

$$AIC = -2\ell(\hat{\beta},\hat{\sigma}^2) + 2p^* = n\ln 2\pi + n\ln\hat{\sigma}^2 + \frac{(y-X\hat{\beta})^T(y-X\hat{\beta})}{\hat{\sigma}^2} + 2p^*$$

ale $\hat{\sigma}^2 = \frac{1}{n}(y-X\hat{\beta})^T(y-X\hat{\beta}) = \frac{SSE}{n}$, tedy

$$AIC = \underbrace{n\ln 2\pi + n}_{C} + n\ln\frac{SSE}{n} + 2p^*$$

alternativně: $$AIC = n\ln\frac{SSE}{n} + 2p^*$$

POZN:
- hledáme model s minimální hodnotou AIC
- AIC není míra kvality modelu, je užitečná pro porovnávání modelů.

AIC v R: AIC(·) počítá $n\ln 2\pi + n + n\ln\frac{SSE}{n} + 2p^*$

kde p^* je počet par. β, σ^2 (včetně interceptu)

extractAIC(·) počítá $n\ln\frac{SSE}{n} + 2p^*$

p^* je jen počet par. β (včetně interceptu)

- (Schwarzovo) bayesovské informační kritérium BIC

$$BIC = -2\ell(\hat{\theta}) + p^* \ln n$$

více penalizuje počet parametrů ⟹ vybírá jednodušší modely s jednodušší interpretací než AIC

BIC vyžaduje významnější příspěvek proměnné, aby byla zařazena do modelu

(AIC se zaměřuje na predikci, BIC je kompromis mezi interpretovatelností a predikcí)

R: BIC(·), nebo AIC(·), extractAIC(·) s volbou

$k = \log(\text{nobs}(\text{fit}))$

Př. data HALD

E) PRESS statistika

pokud je pro nás důležitá kvalita predikce, lze použít pro ~~srovn~~ srovnání modelů statistiku

$$PRESS = \sum_{i=1}^{n} \hat{e}_{(-i)}^2 = \sum_{i=1}^{n} \left(\frac{\hat{e}_i}{1-h_{ii}}\right)^2$$

vybírá se model s minimální hodnotou této statistiky

Metody výběru modelu

1) Vyhodnocení všech možných modelů

pro T dostupných regresorů tzn. nafitovat 2^T modelů

a pak je porovnat pomocí nějakého kritéria

- náročné pro velké T (např. T=20 znamená 1024 modelů!)

2) Zpětná eliminace (backward elimination)

začneme s plným modelem a v každém kroku odstraníme jednu proměnnou, která nejméně přispívá modelu (měřeno F stat) nebo jejíž odstranění znamená největší zlepšení modelu (měřeno AIC)

algoritmus:

1) naložíme model se všemi proměnnými
2) pro každou proměnnou spočteme částečnou F statistiku (nebo t-statistiku) jako by právě byla přidána do modelu
(tzn. se předpokládá, že ostatní proměnné tam jsou)
3) pokud je nejnižší F statistika menší než kritická hodnota F_{OUT}, vyřadíme z modelu proměnnou s nejnižší hodnotou F
($F_{OUT} = F_{1-\alpha_{OUT}}(1, n-p)$, kde p je aktuální počet regresorů v modelu, včetně interceptu, $\alpha_{OUT} = 0{,}05; 0{,}1; \ldots$)
4) opakujeme kroky 2) a 3) dokud všechny částečné F statistiky nejsou větší než příslušná kritická hodnota F_{OUT}
(tzn. nelze vyřadit žádnou proměnnou)

Pozn. místo F lze používat AIC

3) Dopředná regrese (forward regression)

začneme pouze s interceptem (nebo nějakým minimálním modelem) a v každém kroku přidáme jednu proměnnou, která má za následek největší zlepšení modelu
(největší zvýšení F nebo největší pokles AIC)

- tato metoda neumožňuje odstranit proměnnou, která již do modelu byla jednou přidána algoritmus:

1) nalezneme minimální model
2) pro každou dostupnou proměnnou spočteme F statistiku pro test významnosti jejího přidání do modelu
3) pokud některá z těchto F statistik překračuje kritickou hodnotu F_{in}, přidáme do modelu proměnnou s nejvyšší hodnotou F stat.
4) opakujeme kroky 2) a 3) dokud všechny F-stat. pro zbývající proměnné nebudou menší než F_{in} nebo dokud nezbyde žádná proměnná na přidání do modelu

POZN.: i když tento postup zjednodušuje výběr modelu, často bohužel vede na zařazení proměnných, které nemají významný příspěvek, jakmile jsou zařazeny další proměnné

4) Postupná regrese (stepwise regression)

kombinace dvou předchozích metod, v každém kroku algoritmu přidáme jednu proměnnou a poté zkontrolujeme, zda není možné nějakou odebrat

budeme potřebovat dvě kritické hodnoty F_{in}, F_{out} - pro použití F statistiky

algoritmus:

1) nalezneme minimální model
2) zjistíme, zda přidání ~~...~~ nějaké další proměnné může zlepšit model (F nebo AIC), pokud ano přidáme do modelu proměnnou, která má za následek největší zlepšení modelu (největší pokles AIC)
3) v novém modelu zjistíme, zda nelze některou proměnnou

vynechat (opět pomocí AIC nebo F), pokud ano, vynecháme proměnnou, jejíž vyřazení má za následek největší zlepšení modelu (největší pokles AIC)

4) opakujeme kroky 2) a 3) do té doby, až nebude možné přidat ~~nebo~~ ani ubrat žádnou proměnnou

POZN: Princip marginality

- pokud jsou v modelu vyšší mocniny nějakého regresoru, měly by tam být obsaženy i všechny jeho nižší mocniny (i když jsou případně nevýznamné)
- pokud je v modelu obsažena interakce dvou regresorů, měly ~~být~~ by tam být i oba individuální regresory
- s každou interakcí vyššího řádu, by měl model obsahovat i všechny interakce řádu nižšího (a:b:c → a:b, a:c, b:c)

POZN: jakmile nalezneme optimální model, je třeba řádně ověřit jeho adekvátnost