Převzorkování a regularizace

Matematické metody pro ITS (11MAMY)

Jan Přikryl

s využitím díla G. James et al., *An Introduction to Statistical Learning*

10. přednáška 11MAMY

úterý 21. března 2023

verze: 2023-03-20 16:17

Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní

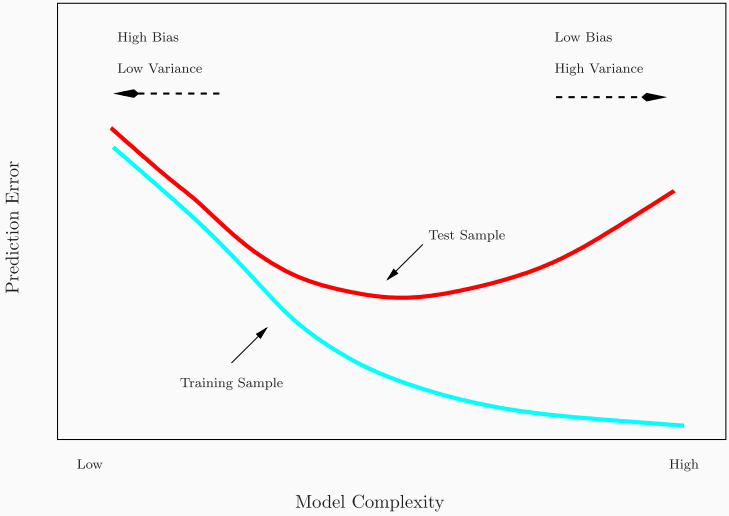
Křížová validace

- V této části se seznámíme s metodami **převzorkování**
- Zaměříme se na *křížovou validaci*, vynecháme *bootstrap*.
- Tyto metody opětovně prokládají model našeho zájmu vzorky vytvářenými z trénovacího souboru s cílem získat dodatečné informace o proloženém modelu.
- Poskytují například odhady chyby předpovědi na testovacím souboru (CV) a směrodatnou odchylku a zkreslení našich odhadů parametrů (bootstrap).

Trénovací chyba versus testovací chyba

- Připomeňte si rozdíl mezi **trénovací chybou** a **testovací chybou**.
- **Testovací chyba** je průměrná chyba, která vzniká při použití metody statistického učení k predikci odpovědi na novém pozorování, takovém, které se nepoužilo při tréninku metody.
- Naproti tomu **trénovací chyba** se dá snadno vypočítat tak, že metodu statistického učení aplikujeme na pozorování použitá k jejímu tréninku.
- Ale míra trénovací chyby je často zcela odlišná od míry testovací chyby a především může ta první z nich dramaticky podhodnocovat tu druhou.

Efektivita: trénovací soubor versus testovací soubor



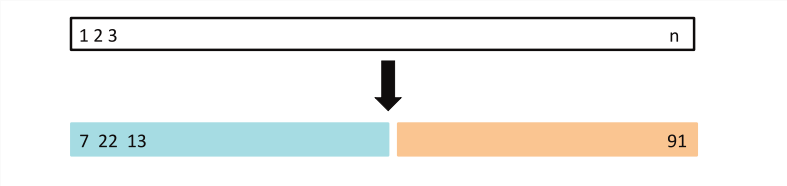
Více o odhadech chyby předpovědi

- Nejlepší řešení: velký k tomu určený soubor dat. Často není k dispozici.
- Některé metody provádějí **matematickou úpravu** míry trénovací chyby s cílem odhadnout míru testovací chyby. Patří k nim **Cp statistika**, **AIC** a **BIC**. Mluví se o nich jinde v tomto kurzu.
- Zde se místo toho zabýváme třídou metod, které odhadují testovací chybu tak, že **odloží stranou** z procesu prokládání podmnožinu trénovacích pozorování a pak aplikují metodu statistického učení na tato odložená pozorování.

Přístup používající validační soubor

- Zde náhodně rozdělíme dostupný soubor vzorků na dvě části: **tréninkovou sadu** a **validační** neboli **odloženou sadu**.
- Model se proloží na tréninkové sadě a proložený model se použije k předpovědi odpovědí na pozorování ve validační sadě.
- Výsledná chyba na validační sadě nám dává odhad testovací chyby. Ta se obvykle posuzuje pomocí MSE v případě kvantitativní odpovědí a pomocí míry chybné klasifikace v případě kvalitativní (diskrétní) odpovědi.

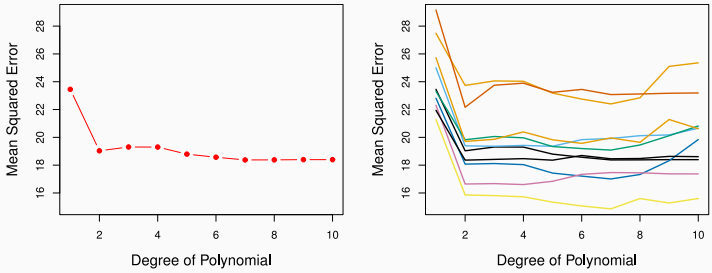
Proces validace



Náhodné rozdělení na dvě poloviny: levá část je trénovací sada, pravá část je validační sada.

Příklad: údaje o automobilech

- Chceme porovnat lineární členy s členy vyšších řádů v polynomech užitých v lineární regresi.
- Náhodně rozdělíme 392 pozorování na dvě sady, trénovací sadu se 196 datovými body a validační sadu obsahující zbylých 196 pozorování.



Levý panel ukazuje jediné rozdělení; pravý panel ukazuje více různých rozdělení.

Nevýhody přístupu s validačním souborem

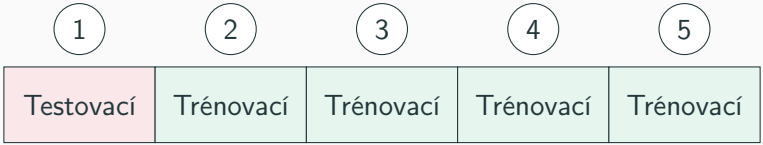
- Validační odhad testovací chyby může být vysoce proměnlivý, závisí totiž přesně na tom, která pozorování se zahrnou do tréninkové sady a která jsou zahrnuta do validační sady.
- U validačního přístupu se k proložení modelu používá pouze podmnožina pozorování – ta, která jsou zahrnuta do tréninkové sady a ne ta ve validační sadě.
- To napovídá, že chyba na validační sadě může mít tendenci **nadhodnocovat** testovací chybu pro model proložený celým souborem dat. **Q:** Proč?

K-násobná křížová validace

- *Široce používaný přístup* k odhadování testovací chyby.
- Odhady se dají použít k výběru nejlepšího modelu a k získání představy o testovací chybě finálního zvoleného modelu.
- Myšlenka zde je náhodně rozdělit data do K stejně velkých částí. Vynecháme část k, proložíme model zbylými $K - 1$ částmi (kombinovaně) a pak získáme předpovědi pro odloženou k-tou část.
- Toto se provádí po řadě pro každou část $k = 1, 2, \ldots, K$ a pak se výsledky zkombinují.

K-násobná křížová validace podrobně

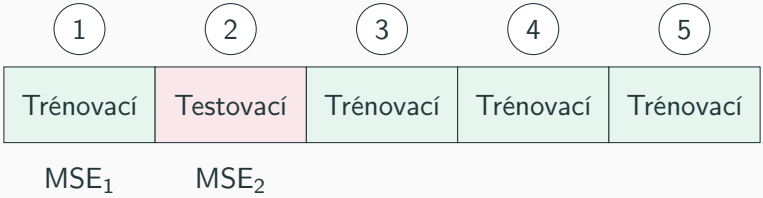
Rozdělíme data do K zhruba stejně velkých částí (zde je $K = 5$).



MSE_1

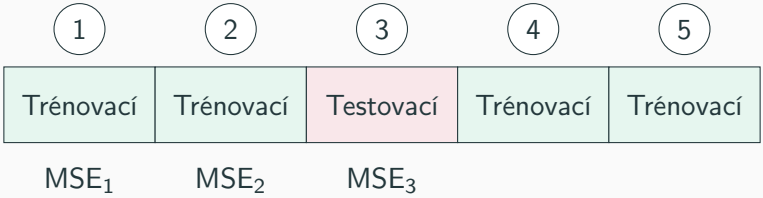
K-násobná křížová validace podrobně

Rozdělíme data do K zhruba stejně velkých částí (zde je $K = 5$).



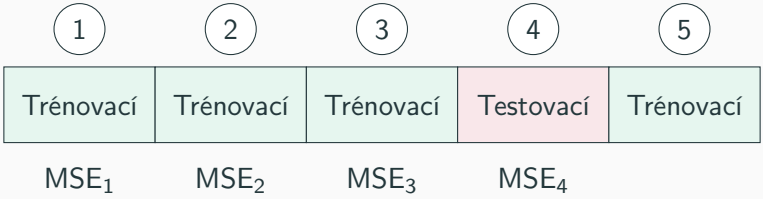
K-násobná křížová validace podrobně

Rozdělíme data do K zhruba stejně velkých částí (zde je $K = 5$).



K-násobná křížová validace podrobně

Rozdělíme data do K zhruba stejně velkých částí (zde je $K = 5$).



K-násobná křížová validace podrobně

Rozdělíme data do K zhruba stejně velkých částí (zde je $K = 5$).

K-násobná křížová validace podrobně

Rozdělíme data do K zhruba stejně velkých částí (zde je $K = 5$).

Výsledný odhad testovací MSE bude vážený průměr jednotlivých MSE_k.

- Nechť těch K částí jsou $C_1, C_2, \ldots, C_K$, kde C_k označuje indexy pozorování v části k. V části k je n_k pozorování; pokud n je násobkem K, je $n_k = n/K$.
- Vypočítáme

$$\mathrm{CV}_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{MSE}_k,$$

  kde $\mathrm{MSE}_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$ a $\hat{y}_i = f_k(\mathbf{x}_i)$ je výstup pro pozorování i získaný modelem f_k natrénovaným z dat s odloženou částí k.
- Položíme-li $K = n$, je výsledkem n-násobná validace neboli **křížová validace s vynecháním jednoho** (LOOCV, angl. *leave-one out cross-validation*).

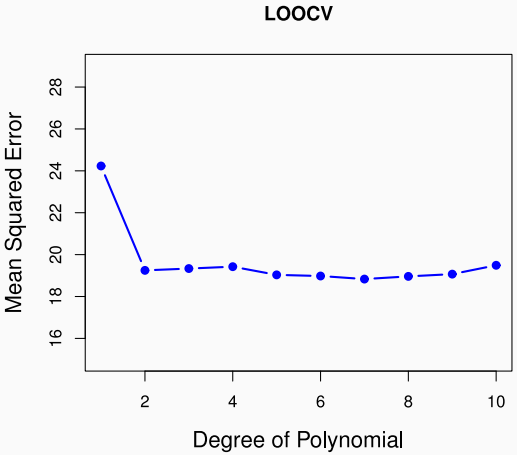
Pěkný speciální případ!

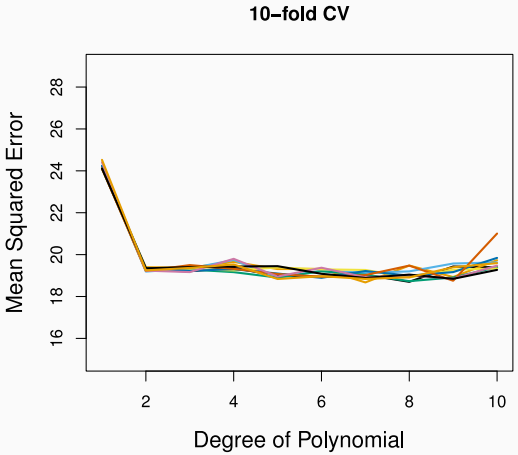
- U lineární nebo polynomiální regrese metodou nejmenších čtverců je tu překvapivý trik, díky němuž je cena LOOCV stejná, jako je cena proložení jediným modelem. Platí následující vzorec:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

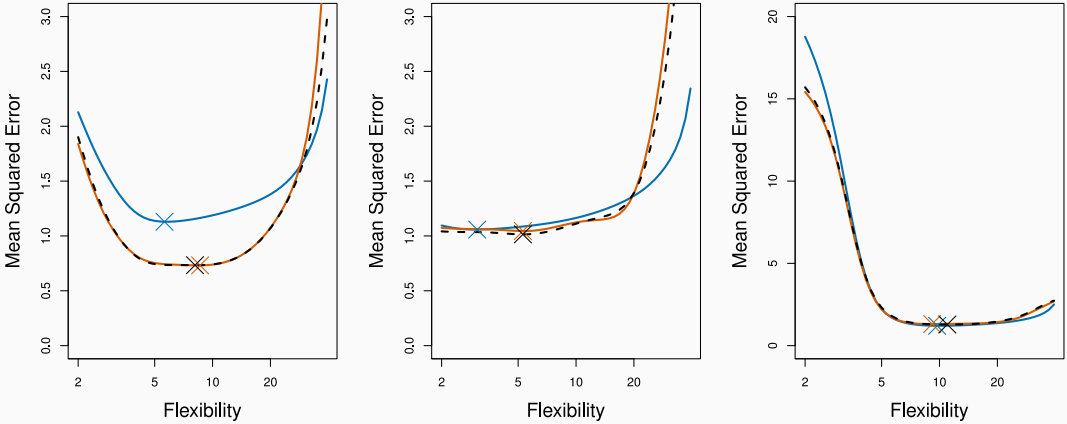
  kde $\hat{y}_i$ je i-tá proložená hodnota z původní aproximace metodou nejmenších čtverců a h_i je účinek (diagonální prvek „stříškové“ matice; podrobnosti jsou v knize). Je to jako obvyklá MSE s tou výjimkou, že i-té reziduum je děleno $1 - h_i$.
- LOOCV je někdy užitečná, ale typicky *neprotřese* data dostatečně. Odhady z jednotlivých složek jsou vysoce *korelovány* a jejich průměr má tudíž *vysoký rozptyl*.
- Lepší volba je $K = 5$ nebo 10.

Data o automobilech zkoumaná znovu





Skutečná a odhadnutá testovací MSE pro simulovaná data



Další problémy s křížovou validací

- Jelikož každá trénovací sada je pouze $(K-1)/K$-krát tak velká jako původní trénovací sada, odhady chyby předpovědi budou typicky zkresleny směrem nahoru. **Q:** Proč?
- Toto zkreslení se minimalizuje při $K = n$ (LOOCV), ale tento odhad má vysoký rozptyl, jek jsme uvedli dříve.
- $K = 5$ nebo 10 dává dobrý kompromis pro vyvážení tohoto vztahu zkreslení a rozptylu.

Křížová validace pro klasifikační úlohy

- Rozdělíme data na K zhruba stejně velikých částí $C_1, C_2, \ldots, C_K$. C_k označuje indexy pozorování v části k. V části k je n_k pozorování: jestliže n je násobkem K, pak $n_k = n/K$.
- Vypočítáme

$$\mathrm{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{Err}_k$$

  kde $\mathrm{Err}_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.
- Odhadnutá směrodatná odchylka CV_K je

$$\widehat{\mathrm{SE}}(\mathrm{CV}_K) = \sqrt{\sum_{k=1}^{K} (\mathrm{Err}_k - \overline{\mathrm{Err}_k})^2/(K-1)}$$

- Toto je užitečný odhad, ale, přesněji řečeno, není zcela správně. **Q:** Proč?

Křížová validace: správně a chybně

- Uvažujme jednoduchý klasifikátor aplikovaný na nějaká dvoutřídní data:
  a) Začínáme s 5000 prediktory a 50 vzorky a najdeme těch 100 prediktorů, které mají největší korelaci se štítky tříd.
  b) Pak použijeme nějaký klasifikátor, jako je třeba logistická regrese, pouze na těchto 100 prediktorů.

Jak odhadneme účinnost tohoto klasifikátoru na testovacím souboru?

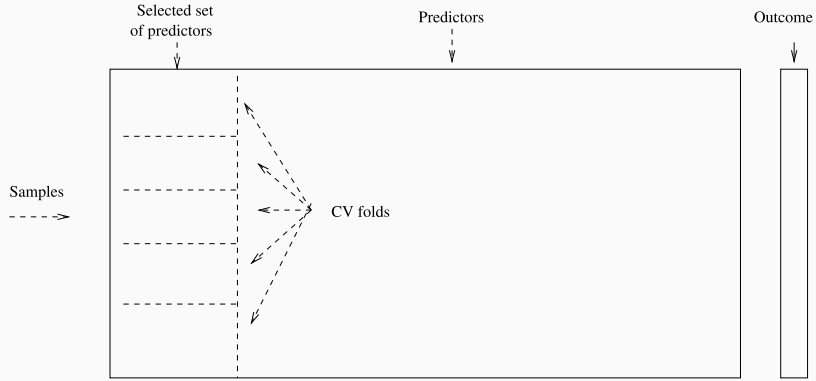
Můžeme použít křížovou validaci v kroku 2 a zapomenout na krok 1?

- To by ignorovalo skutečnost, že naše procedura v kroku 1 *již viděla štítky trénovacích dat* a zužitkovala je. To je forma tréninku a musí to být do validačního procesu zahrnuto.
- Je snadné nasimulovat realistická data se štítky tříd nezávisejícími na výstupu, takže skutečná testovací chyba je 50 %, ale odhad chyby křížovou validací, který ignoruje krok 1, bude nula! (Zkuste to udělat sami.)
- Tuto chybu dělá mnoho autorů ve článcích i ve vysoce profilovaných časopisech například z genomiky.

Chybný a správný způsob

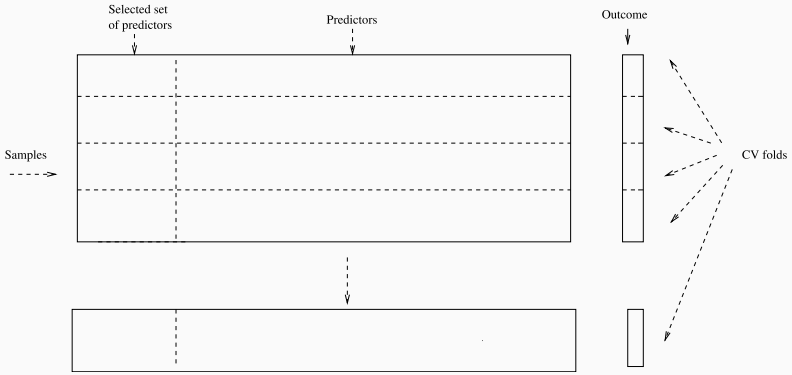
- *Chybně:* Použijeme křížovou validaci v kroku 2.
- *Správně:* Použijeme křížovou validaci v krocích 1 a 2.

Chybný způsob



Texty: Vzorky, Vybraná sada prediktorů, Složky křížové validace, Prediktory, Výsledky

Správný způsob



Texty: Vzorky, Vybraná sada prediktorů, Prediktory, Výsledky, Složky křížové validace

Výběr optimálního modelu
Regularizace

- Připomeňme si lineární model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

- V přednáškách, které následují, se budeme zabývat některými přístupy, které rozšiřují rámec lineárního modelu. V přednáškách, které pokrývají kapitolu 7 učebnice, zobecňujeme lineární model tak, aby zahrnul **nelineární**, ale **aditivní** vztahy.
- V přednáškách pokrývajících kapitolu 8 se zabýváme ještě obecnějšími **nelineárními** modely.

Chválíme lineární modely!

- Nehledě na svou jednoduchost má lineární model zřetelné výhody co do své **interpretovatelnosti** a často vykazuje dobré **výsledky v předpovědích**.
- V této přednášce se tudíž budeme zabývat některými způsoby, jimiž lze jednoduchý lineární model vylepšit tak, že zaměníme obvyklé prokládání nejmenšími čtverci nějakým alternativním způsobem aproximace.

Proč se zabývat alternativami k nejmenším čtvercům?

- **Přesnost předpovědi:** Zejména při $p > n$, k regulaci rozptylu.
- **Interpretovatelnost modelu:** Odstraněním nepodstatných vlastností — to jest tím, že položíme odpovídající odhady koeficientů rovny nule — můžeme získat model, který se snáze interpretuje. Uvedeme některé přístupy k automatickému provádění **volby vlastností.**

Tři třídy metod

- **Výběr podmnožiny.** Identifikujeme podmnožinu z těch p prediktorů, o níž soudíme, že má vztah k odpovědi. Pak proložíme nejmenšími čtverci model tou redukovanou množinou proměnných.
- **Smršťování.** Proložíme model zahrnující všech p prediktorů, ale odhadnuté koeficienty srazíme směrem k nule vzhledem k odhadům nejmenších čtverců. Toto smrštění (známé také jako **regularizace**) má efekt ve snížení rozptylu a může také provádět výběr proměnných.
- **Dimenzionální redukce.** Promítneme těch p prediktorů na M-rozměrný podprostor, kde $M < p$. Toho se dosáhne tak, že vypočítáme M různých **lineárních kombinací**, neboli **projekcí** těch proměnných. Pak se těchto M projekcí použije jako prediktory k proložení lineárního regresního modelu nejmenšími čtverci.

Algoritmy pro model s výběrem nejlepší podmnožiny a postupný výběr modelu

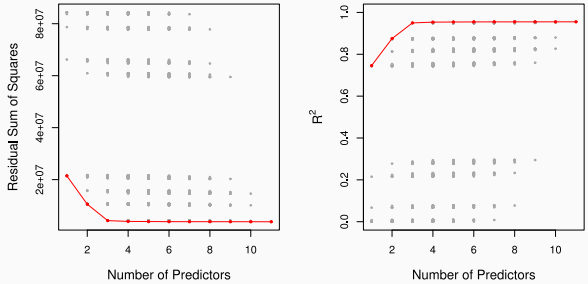
Výběr nejlepší podmnožiny

a) Označme $\mathcal{M}_0$ **nulový model**, který neobsahuje žádné prediktory. Tento model pro každé pozorování prostě předpovídá střední hodnotu vzorku.

b) Pro $k = 1, 2, \ldots, p$:
   (a) Prolož všech $\binom{p}{k}$ modelů, které obsahují přesně k prediktorů.
   (b) Vyber z těchto $\binom{p}{k}$ modelů ten nejlepší a označ jej $\mathcal{M}_k$. **Nejlepší** se zde definuje jako mající nejmenší RSS nebo ekvivalentně největší R^2.

c) Vyber jediný nejlepší model z modelů $\mathcal{M}_0, \ldots, \mathcal{M}_p$ na základě chyby křížové validace, C_p (AIC), BIC nebo upraveného R^2.

Příklad: soubor kreditních dat



Pro každý možný model obsahující podmnožinu deseti prediktorů v datovém souboru `Credit` *jsou zde znázorněny RSS a* R^2*. Červené ohraničení sleduje* ***nejlepší*** *model pro daný počet prediktorů podle RSS a* R^2*. Ačkoli soubor dat obsahuje pouze deset prediktorů, osa x má rozsah od 1 do 11, neboť jedna z proměnných je kategoriální a nabývá tří hodnot, což vede k vytvoření dvou fiktivních proměnných.*

Rozšíření na jiné modely

- Ačkoli jsme zde výběr nejlepší podmnožiny ukázali pro regresi nejmenšími čtverci, stejné myšlenky se vztahují i na jiné typy modelů, jako je logistická regrese.
- **Deviance** — záporně vzatý dvojnásobek maximalizované logaritmické věrohodnosti — hraje roli RSS pro širší třídu modelů.

Postupný výběr

- Výběr nejlepší podmnožiny se z výpočtových důvodů nedá použít pro velmi velká p. **Proč ne?**
- Když p je velké, může výběr nejlepší podmnožiny také trpět statistickými problémy: čím větší prostor pro vyhledávání, tím větší je šance najít modely, které na tréninkových datech vypadají dobře, i když na budoucích datech nemusejí mít žádnou vypovídací hodnotu.
- Enormní prostor pro vyhledávání tudíž může vést k **přeurčení** a vysokému rozptylu odhadů koeficientů.
- Z obou těchto důvodů jsou atraktivními alternativami k výběru nejlepší podmnožiny metody **postupného výběru**, které zkoumají mnohem omezenější soubor modelů.

Postupná dopředná selekce

- Postupná dopředná selekce začíná modelem, který neobsahuje žádné prediktory, a pak k modelu prediktory přidává jeden po druhém, dokud v modelu nejsou všechny prediktory.
- Konkrétně se v každém kroku k modelu přidává proměnná, která prokládané aproximaci dává největší **dodatečné** zlepšení.

Postupná dopředná selekce

a) Označme $\mathcal{M}_0$ **nulový** model, který neobsahuje žádné prediktory.

b) Pro $k = 0, \ldots, p-1$:

   2.1 Uvažujme všech $p - k$ modelů, které rozšiřují prediktory v $\mathcal{M}_k$ o jeden prediktor navíc.

   2.2 Vyberme **nejlepší** z těchto $p - k$ modelů a nazvěme jej $\mathcal{M}_{k+1}$. **Nejlepší** zde znamená, že model má nejmenší RSS nebo největší R^2.

c) Vybereme jediný nejlepší model z modelů $\mathcal{M}_0, \ldots, \mathcal{M}_p$ na základě chyby předpovědi křížové validace, C_p (AIC), BIC nebo upraveného R^2.

Více o postupné dopředné selekci

- Výpočetní výhoda před výběrem nejlepší podmnožiny je jasná.
- Není zaručeno, že se najde ten nejlepší model ze všech 2^p modelů obsahujících podmnožiny daných p prediktorů. **Proč ne? Uveďte příklad.**

Příklad s kreditními daty

Počet prom.	Nejlepší podmnožina	Postupná dopředná selekce
Jedna	`rating`	`rating`
Dvě	`rating`, `income`	`rating`, `income`
Tři	`rating`, `income`, `student`	`rating`, `income`, `student`
Čtyři	`cards`, `income`	`rating`, `income`
	`student`, `limit`	`student`, `limit`

První čtyři vybrané modely pro výběr nejlepší podmnožiny a postupnou dopřednou selekci na souboru dat `Credit`*. První tři modely jsou identické, ale čtvrté modely se liší.*

Postupná zpětná eliminace

- Podobně jako postupná dopředná selekce představuje **postupná zpětná eliminace** efektivní alternativu k výběru nejlepší podmnožiny.
- Avšak na rozdíl od postupné dopředné selekce začíná úplným modelem nejmenších čtverců obsahujícím všech p prediktorů a pak iterativně odstraňuje jeden po druhém nejméně užitečné prediktory.

Postupná zpětná eliminace

a) Označme $\mathcal{M}_p$ **úplný** model, který obsahuje všech p prediktorů.

b) Pro $k = p, p-1, \ldots, 1$:

   2.1 Uvažujme všech k modelů, které obsahují všechny prediktory z $\mathcal{M}_k$ kromě jednoho, takže mají celkem $k-1$ prediktorů.

   2.2 Vybereme z těchto k modelů ten **nejlepší** a označíme jej $\mathcal{M}_{k-1}$. **Nejlepším** je zde míněn model s nejmenším RSS nebo největším R^2.

c) Vybereme jediný nejlepší model z modelů $\mathcal{M}_0, \ldots, \mathcal{M}_p$ na základě chyby předpovědi křížové validace, C_p (AIC), BIC nebo upraveného R^2.

Více o postupné zpětné eliminaci

- Podobně jako postupná dopředná selekce prochází postupná zpětná eliminace pouze $1 + p(p+1)/2$ modelů, a tak může být použita v situacích, kdy p je pro použití výběru nejlepší podmnožiny příliš velké.
- Podobně jako postupná dopředná selekce nezaručuje postupná zpětná eliminace, že poskytne **nejlepší** model zahrnující některou podmnožinu daných p prediktorů.
- Zpětná eliminace vyžaduje, aby **počet vzorků n byl větší než počet proměnných p** (takže lze proložit úplný model). Naproti tomu dopředná selekce může být použita dokonce když $n < p$, a tak je to jediná životaschopná metoda založená na podmnožinách v případě, že p je velmi velké.

Výběr optimálního modelu

- Model obsahující všechny prediktory bude vždy mít nejmenší RSS a největší R^2, neboť tyto veličiny se vztahují k trénovací chybě.
- Přejeme si zvolit model s nízkou testovací chybou, ne model s nízkou trénovací chybou. Připomínáme, že trénovací chyba je obvykle špatným odhadem testovací chyby.
- V důsledku toho nejsou RSS a R^2 vhodné pro výběr nejlepšího modelu z kolekce modelů s různými počty prediktorů.

Odhad testovací chyby

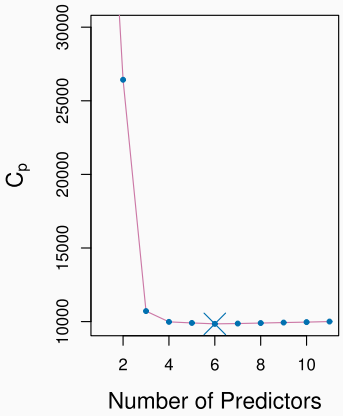
Odhadování testovací chyby: dva přístupy

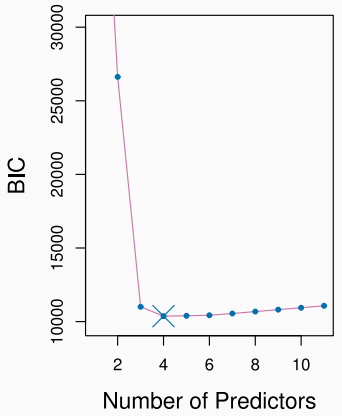
- Můžeme testovací chybu odhadnout nepřímo tak, že provedeme *úpravu trénovací chyby*, která vezme v úvahu zkreslení působené přeurčováním.
- Můžeme testovací chybu odhadnout přímo tak, jak jsme probírali v předchozích přednáškách,
  - buď použitím přístupu s *validačním souborem*
  - nebo použitím *křížové validace*
- Oba přístupy budeme ilustrovat v dalším.

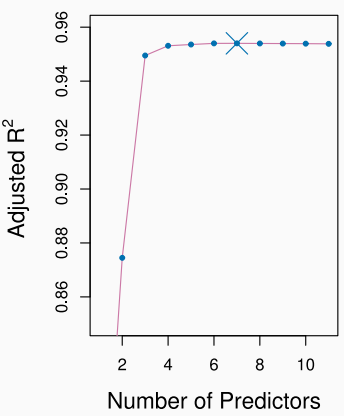
C_p, AIC, BIC a upravené R^2

- Tyto postupy přizpůsobují trénovací chybu velikosti modelu a mohou se použít k výběru z množiny modelů s různými počty proměnných.
- Následující obrázek znázorňuje C_p, BIC a upravené R^2 pro nejlepší model každé velikosti získaný na souboru dat `Credit` výběrem nejlepší podmnožiny.

Příklad s kreditními daty







Mallowsovo C_p:

$$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2),$$

kde d je celkový počet použitých parametrů a $\hat{\sigma}^2$ je odhad rozptylu chyby ϵ spojené s měřením každé odpovědi.

- Optimální model podle tohoto kritéria je kompromis ovlivněný velikostí vzorku, velikostí efektů různých prediktorů a stupněm kolinearity mezi nimi.
- Původně navrženo jako kritérium pro výběr z mnoha alternativních podmnožin regresorů.

Akakikeho informační kritérium AIC je definováno pro velkou třídu prokládaných modelů na základě jejich maximální věrohodnosti:

$$\mathrm{AIC} = -2 \log \hat{L} + 2 \cdot d,$$

kde $\hat{L}$ je maximalizovaná hodnota věrohodnostní funkce pro odhadovaný model.

- AIC odhaduje relativní objem informace ztracený daným modelem: čím méně informace model ztratí, tím vyšší je kvalita modelu.
- V případě lineárního modelu s gaussovskými chybami jsou maximální věrohodnost a nejmenší čtverce stejná věc a C_p a AIC jsou ekvivalentní. **Dokažte to.**

Podrobně o BIC

Při prokládání dat je možné zvýšit věrohodnost modelu přidáním parametrů, ale může to vést k přeurčení. BIC i AIC proto zavádějí penalizační člen pro počet parametrů:

$$\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n) d\hat{\sigma}^2) \approx -2 \log \hat{L} + 2 \cdot d$$

- C_p i BIC nabývají malých hodnot pro modely s nízkou testovací chybou, obecně vybíráme ten model, který má nejmenší BIC.
- BIC nahrazuje $2d\hat{\sigma}^2$ v C_p členem $\log(n)d\hat{\sigma}^2$, kde n je počet pozorování.
- Jelikož pro jakékoli $n > 7$ je $\log n > 2$, umísťuje statistika BIC obecně těžší penaltu na modely s mnoha proměnnými, a tudíž vybírá menší modely než C_p. Viz obrázek na slajdu 43.

Upravené R^2

Pro model nejmenších čtverců s d proměnnými se upravená statistika R^2 vypočítá jako

$$\text{Upravené } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)},$$

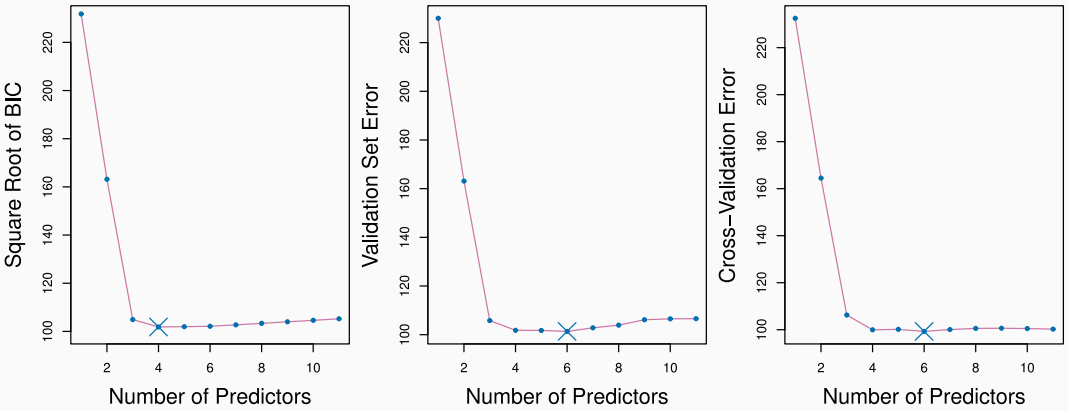
kde TSS je celkový součet čtverců.

- Na rozdíl od C_p, AIC a BIC, kde malá hodnota označuje model s nízkou testovací chybou, velká hodnota upraveného R^2 označuje model s malou testovací chybou.
- Maximalizace upraveného R^2 je ekvivalentní minimalizaci $\text{RSS}/(n-d-1)$. Zatímco RSS s růstem počtu proměnných v modelu vždy klesá, $\text{RSS}/(n-d-1)$ může růst nebo klesat, a to díky přítomnosti d ve jmenovateli.
- Na rozdíl od statistiky R^2 se v upravené statistice R^2 za zahrnutí nepotřebných proměnných do modelu *platí cena*. Viz obrázek na slajdu 43.

Validace a křížová validace

- Každá z metod výběru regresorů vrací posloupnost modelů $\mathcal{M}_k$ indexovaných velikostí modelu $k = 0, 1, 2, \ldots$. Naším úkolem zde vybrat optimální $\hat{k}$. Jakmile je vybereme, vracíme model $\mathcal{M}_{\hat{k}}$.
- Vypočítáme chybu na validační množině nebo chybu křížové validace pro každý uvažovaný model $\mathcal{M}_k$ a pak vybereme k, pro něž je výsledná odhadnutá testovací chyba nejmenší.
- Tento postup má vůči AIC, BIC, C_p a upravenému R^2 tu výhodu, že poskytuje přímý odhad testovací chyby a *nevyžaduje odhad rozptylu chyby σ^2.*
- Dá se také použít v širším rozmezí úloh s výběrem modelu, dokonce v případech, kdy je obtížné stanovit počet stupňů volnosti v modelu (tj. počet prediktorů v modelu) nebo je obtížné odhadnout rozptyl chyby σ^2.

Příklad s kreditními daty



Podrobněji k předchozímu obrázku

- Validační chyby byly vypočítány náhodným výběrem tří čtvrtin pozorování za trénovací sadu a zbytku jako validační množinu.
- Křížová validace byla počítána s $k = 10$ složkami. V tomto případě metoda validace i metoda křížové validace obě dávaly jako výsledek model s šesti proměnnými.
- Nicméně všechny tři přístupy naznačují, že modely se čtyřmi, pěti a šesti proměnnými jsou co do svých testovacích chyb zhruba ekvivalentní.

V této situaci volíme model pomocí **pravidla jedné směrodatné chyby**: Vypočítáme nejprve směrodatnou odchylku odhadnuté testovací MSE pro každou velikost modelu a pak zvolíme ten nejmenší model, pro nějž leží odhadnutá testovací chyba v rozmezí jedné směrodatné odchylky od nejnižšího bodu na křivce. **Čím se to dá zdůvodnit?**

Regularizace (smršťování)

Hřebenová regrese a metoda Lasso

- Metody výběru podmnožiny používají nejmenší čtverce k prokládání lineárního modelu, který obsahuje podmnožinu prediktorů.
- Jako alternativu můžeme proložit model obsahující všech p prediktorů pomocí techniky, která omezuje nebo regularizuje odhady koeficientů, nebo ekvivalentně, která smršťuje odhady koeficientů směrem k nule.
- Není možná bezprostředně zřejmé, proč by takové omezení mělo prokládanou aproximaci vylepšit, ale ukazuje se, že smrštění odhadů koeficientů může významně snížit jejich rozptyl.

Hřebenová regrese

- Připomínáme, že metoda prokládání metodou nejmenších čtverců odhaduje $\beta_0, \beta_1, \ldots, \beta_p$ pomocí hodnot, které minimalizují

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2.$$

- Naproti tomu odhady koeficientů hřebenové regrese $\hat{\beta}^R$ jsou hodnoty, které minimalizují

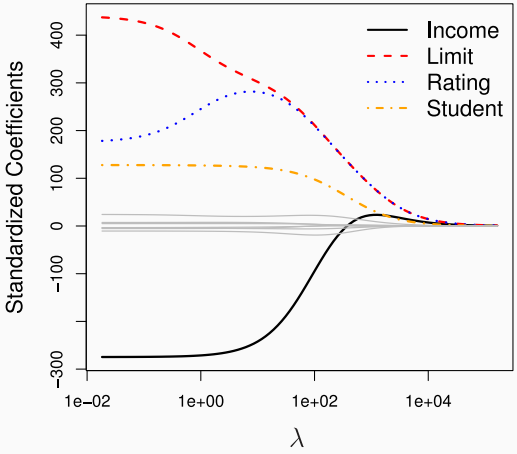
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2,$$

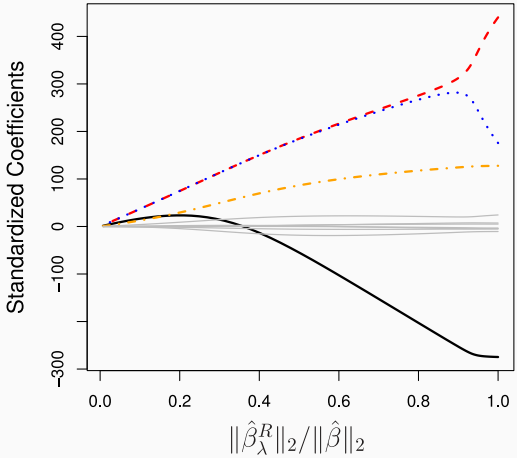
kde $\lambda \geq 0$ je **ladicí parametr**, který je třeba stanovit odděleně.

Hřebenová regrese: pokračování

- Jako u nejmenších čtverců hledá hřebenová regrese odhady koeficientů, které prokládají data dobře, a to tak, že dělá RSS malé.
- Nicméně druhý člen, $\lambda \sum_j \beta_j^2$, nazývaný **smršťovací penalta** je malý, jsou-li $\beta_1, \ldots, \beta_p$ blízké nule, a tak má efekt smršťování odhadů β_j směrem k nule.
- Ladicí parametr λ slouží k řízení relativního vlivu těchto dvou členů na odhady regresních koeficientů.
- Volba dobré hodnoty λ je kritická; *používá se k tomu křížová validace*.

Příklad s kreditními daty





Podrobnosti k předchozímu obrázku

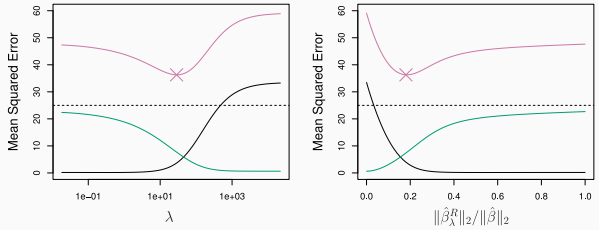
- Na levém panelu odpovídá každá křivka odhadu koeficientu hřebenové regrese pro jednu z deseti proměnných, znázorněnému jako funkce λ.
- Pravý panel zobrazuje stejné odhady hřebenových koeficientů jako levý panel, ale místo toho, abychom na ose x vynášeli λ, vynášíme tam nyní $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$, kde $\hat{\beta}$ označuje vektor odhadů koeficientů nejmenších čtverců.
- Označení $\|\beta\|_2$ znamená L_2 normu vektoru (vyslovuje se to „el dva“), která je definována jako $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$.

Hřebenová regrese: škálování prediktorů

- Standardní odhady koeficientů metody nejmenších čtverců jsou **měřítkově invariantní:** vynásobení x_j konstantou c vede prostě k přeškálování odhadů koeficientů nejmenších čtverců faktorem $1/c$. Jinými slovy, bez ohledu na to, jak je škálován j-tý prediktor, zůstane $\hat{\beta}_j x_j$ beze změny.
- Naproti tomu se odhady koeficientů hřebenové regrese mohou podstatně změnit, vynásobíme-li daný prediktor konstantou, a to díky členu se součtem druhých mocnin koeficient v penalizační části cílové funkce hřebenové regrese.
- Je tudíž nejlepší používat hřebenovou regresi po **normalizaci prediktorů** pomocí vzorce

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

Kompromis mezi zkreslením a rozptylem



Simulovaná data s $n = 50$ pozorováními, $p = 45$ prediktory, všechny s nenulovými koeficienty. Druhá mocnina zkreslení (černě), rozptyl (zeleně), a střední kvadratická testovací chyba (purpurově) pro předpovědi hřebenovou regresí na simulovaném souboru dat jako funkce λ a $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. Vodorovné tečkované čáry označují minimální možnou MSE. Purpurové křížky označují ty modely hřebenové regrese, pro něž je MSE nejmenší.

- Hřebenová regrese má jednu zřejmou nevýhodu: na rozdíl od výběru podmnožiny, který bude obecně vybírat modely zahrnující pouze nějakou podmnožinu proměnných, hřebenová regrese do konečného modelu zahrne všech p prediktorů.
- Metoda **Lasso** je poměrně nedávnou alternativou k hřebenové regresi, která tuto nevýhodu překonává. Koeficienty metody Lasso, $\hat{\beta}_\lambda^L$, minimalizují veličinu

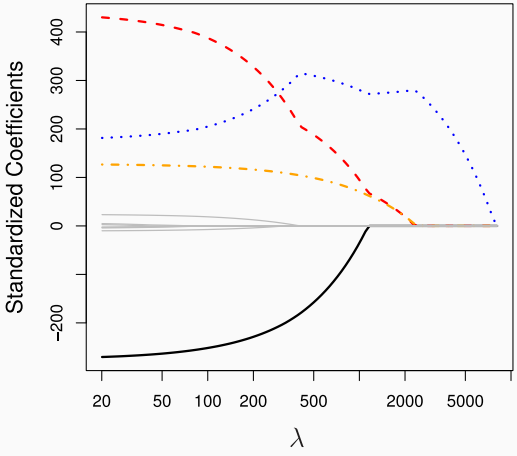
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|.$$

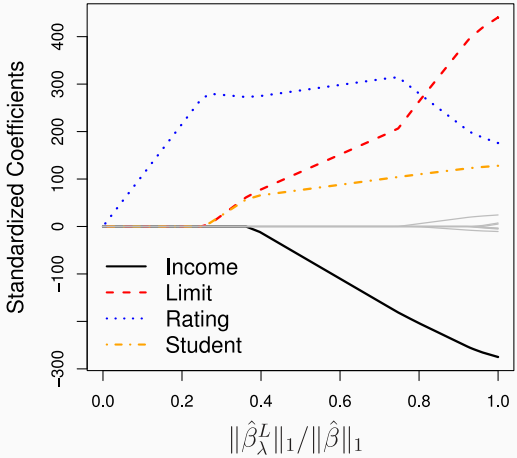
- Ve statistickém žargonu používá metoda Lasso L_1 (vyslovováno jako „el jedna“) penaltu namísto L_2 penalty. Přitom L_1 norma vektoru koeficientů β je dána vztahem $\|\beta\|_1 = \sum|\beta_j|$.

Metoda Lasso: pokračování

- Stejně jako hřebenová regrese smršťuje metoda Lasso odhady koeficientů směrem k nule.
- Avšak v případě metody Lasso má L_1 penalta ten efekt, že pokud je ladicí parametr λ dostatečně velký, donutí to některé z odhadů koeficientů $\hat{\beta}^{\mathsf{L}}_{\lambda}$, aby byly přesně nulové.
- Tudíž velmi podobně jako při výběru nejlepší podmnožiny provádí metoda Lasso **výběr proměnných**.
- Říkáme, že Lasso nám dává **řídké** modely — to jest modely, které zahrnují pouze nějakou podmnožinu proměnných.
- Stejně jako u hřebenové regrese je volba dobré hodnoty λ pro metodu Lasso kritická; metodou volby je opět křížová validace.

Příklad: soubor kreditních dat





Vlastnost výběru proměnných u metody Lasso

Proč je tomu tak, že metoda Lasso, na rozdíl od hřebenové regrese, dává jako výsledek odhady koeficientů, které jsou přesně rovny nule?

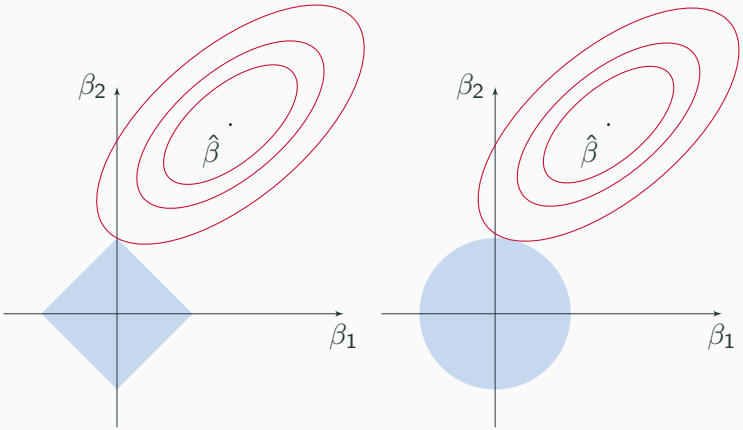
Dá se ukázat, že odhady koeficientů metodou Lasso a hřebenovou regresí řeší po řadě úlohy

$$\underset{\beta}{\text{minimalizuj}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ vzhledem k } \sum_{j=1}^{p} |\beta_j| \leq s$$

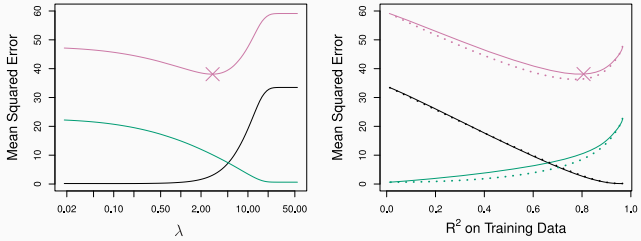
a

$$\underset{\beta}{\text{minimalizuj}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ vzhledem k } \sum_{j=1}^{p} \beta_j^2 \leq s.$$

Metoda Lasso na obrázku

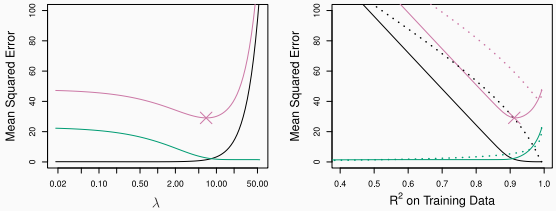


Porovnání metody Lasso a hřebenové regrese



Vlevo: Grafy kvadrátu zkreslení (černá), rozptylu (zelená) a testovací MSE (purpurová) pro metodu Lasso použitou na simulovaném souboru dat ze slajdu 63. Vpravo: Porovnání kvadrátu zkreslení, rozptylu a testovací MSE mezi metodou Lasso (plné čáry) a hřebenovou regresí (čárkovaně). Hodnoty pro obě metody jsou vyneseny proti jejich R^2 na trénovacích datech, což je běžný způsob indexování. Křížky na obou grafech označují model Lasso, pro nějž je MSE nejmenší.

Porovnání metody Lasso a hřebenové regrese: pokračování



Vlevo: Grafy kvadrátu zkreslení (černá), rozptylu (zelená) a testovací MSE (purpurová) pro metodu Lasso. Simulovaná data jsou podobná těm na slajdu 63 až na to, že nyní se k odpovědi vztahují pouze dva prediktory. Vpravo: Porovnání kvadrátu zkreslení, rozptylu a testovací MSE mezi metodou Lasso (plné čáry) a hřebenovou regresí (čárkovaně). Hodnoty pro obě metody jsou vyneseny proti jejich R^2 na trénovacích datech, což je běžný způsob indexování. Křížky na obou grafech označují model Lasso, pro nějž je MSE nejmenší.

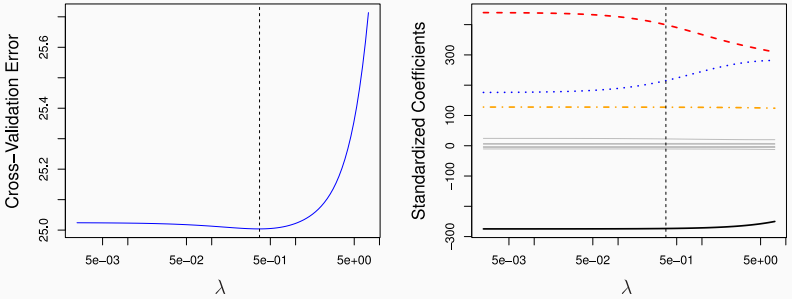
Závěry

- Tyto dva příklady ukazují, že ani hřebenová regrese ani metoda Lasso nepřevládnou univerzálně jedna nad druhou.
- Obecně se dá předpokládat, že Lasso bude pracovat lépe, bude-li odpověď funkcí pouze poměrně malého počtu prediktorů.
- Avšak počet prediktorů, které mají vliv na odpověď, není u reálných souborů dat nikdy znám ă priori.
- K rozhodnutí o tom, který přístup je pro daný soubor dat lepší, se dá použít některá technika typu křížové validace.

Volba parametru λ pro hřebenovou regresi a Lasso

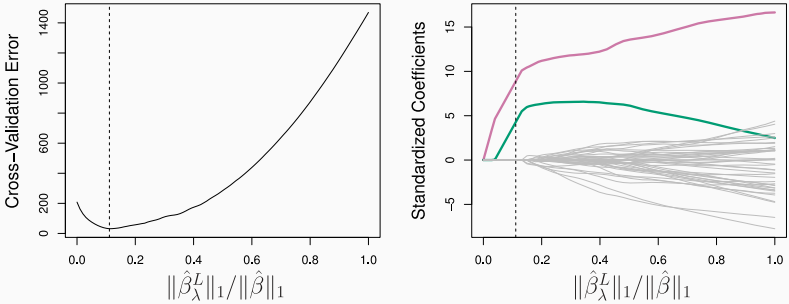
- Stejně jako výběr podmnožiny vyžadují hřebenová regrese a metoda Lasso nějakou metodu ke stanovení toho, který z uvažovaných modelů je nejlepší.
- Potřebujeme tedy metodu pro volbu hodnoty ladicího parametru λ nebo ekvivalentně pro hodnotu omezení s.
- Jednoduchý způsob, jak se vypořádat s tímto problémem nám poskytuje *křížová validace*. Zvolíme si mřížku hodnot λ a vypočítáme míru chyby křížové validace pro každou hodnotu λ.
- Pak zvolíme tu hodnotu ladicího parametru, pro kterou je chyba křížové validace nejmenší.
- Nakonec znovu proložíme model s použitím všech dostupných pozorování a se zvolenou hodnotou ladicího parametru.

Příklad s kreditními daty



Vlevo: Chyby křížové validace vznikající při aplikaci hřebenové regrese na soubor dat `Credit` *s různými hodnotami* λ. *Vpravo:* Odhady koeficientů jako funkce λ. Svislé čárkované linky označují hodnotu λ vybranou křížovou validací.

Příklad se simulovanými daty



Vlevo: MSE u desetisložkové křížové validace pro metodu Lasso aplikovanou na řídký simulovaný soubor dat ze slajdu 38. *Vpravo:* Zde jsou znázorněny příslušné odhady koeficientů metodou Lasso. Svislé čárkované linky označují aproximaci metodou Lasso, pro niž je chyba křížové validace nejmenší.

Redukce počtu parametrů

- Metody, které jsme v této kapitole dosud probírali, spočívaly v prokládání lineárních regresních modelů, pomocí nejmenších čtverců nebo přístupu se smršťováním, při použití původních prediktorů $X_1, X_2, \ldots, X_p$.
- Budeme se nyní zabývat skupinou přístupů, které transformují prediktory a pak nejmenšími čtverci prokládají model užívající transformované proměnné.

Tyto postupy budeme nazývat metodami **dimenzionální redukce**.

Metody dimenzionální redukce: podrobnosti

- Nechť $Z_1, Z_2, \ldots, Z_M$ představují $M < p$ *lineárních kombinací* našich původních p prediktorů. Tedy, pro nějaké konstanty $\phi_{m1}, \ldots, \phi_{mp}$,

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j.$$

- Pomocí obvyklé metody nejmenších čtverců pak prokládáme lineární model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n.$$

- Poznamenáváme, že regresní koeficienty jsou nyní $\theta_0, \theta_1, \ldots, \theta_M$.

Jsou-li konstanty $\phi_{m1}, \ldots, \phi_{mp}$ vybrány vhodně, pak postup dimenzionální redukce může často překonat regresi obvyklými nejmenšími čtverci.

- Všimněme si, že z definice plyne

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij},$$

  kde

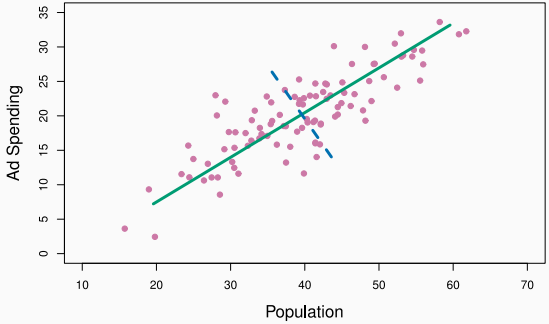
$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj}.$$

- Model tedy můžeme chápat jako speciální případ původního lineárního regresního modelu.
- Dimenzionální redukce slouží k omezení odhadovaných koeficientů β_j, neboť nyní musí koeficienty nabývat tvaru, uvedeného výše.
- Může být přínosem pro kompromis mezi zkreslením a rozptylem.

Regrese hlavních komponent (PCR)

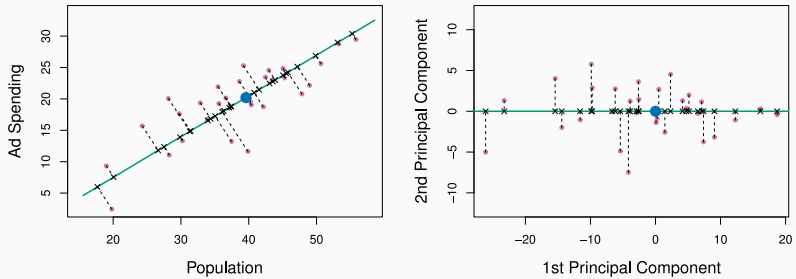
- Zde použijeme **analýzu hlavních komponent** (PCA, angl. *Principal Component Analysis*, probíráno v kapitole 10 ISLR) k zavedení lineárních kombinací prediktorů, které užijeme v naší regresi.
- První hlavní komponenta je ta (normalizovaná) lineární kombinace proměnných, která má největší rozptyl.
- Druhá hlavní komponenta má největší rozptyl s tím omezením, že *není korelována s tou první*.
- A tak dále.
- Při mnoha korelovaných původních proměnných je tedy nahrazujeme malou množinou hlavních komponent zachycující jejich společnou proměnlivost.

Zobrazení PCA



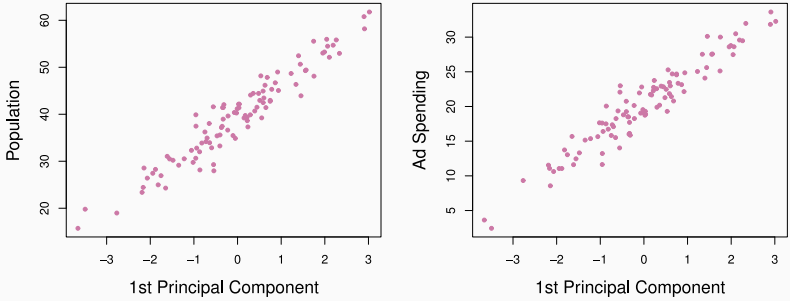
Velikost populace (pop) *a náklady na reklamu* (ad) *pro 100 různých měst jsou zde zobrazeny jako purpurová kolečka. Zelená plná přímka vyznačuje první hlavní komponentu a modrá čárkovaná přímka vyznačuje druhou hlavní komponentu.*

Zobrazení PCA: pokračování



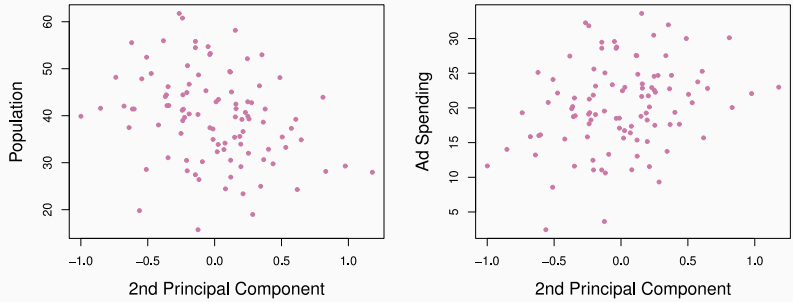
Podmnožina reklamních dat. Vlevo: První hlavní komponenta vybraná tak, aby minimalizovala součet čtverců kolmých vzdáleností od každého bodu, je zobrazena zeleně. Tyto vzdálenosti jsou znázorněny černými čárkovanými úsečkami. Vpravo: Levý panel byl otočen tak, že první hlavní komponenta leží na ose x.

Zobrazení PCA: pokračování



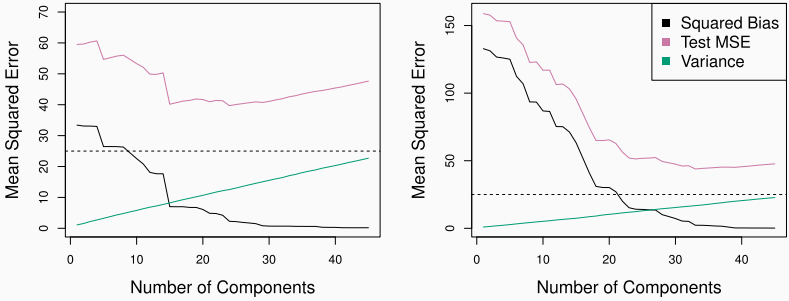
Grafy hodnot první hlavní komponenty z_{i1} *versus* `pop` *a* `ad`. *Vzájemné vztahy jsou silné.*

Zobrazení PCA: pokračování



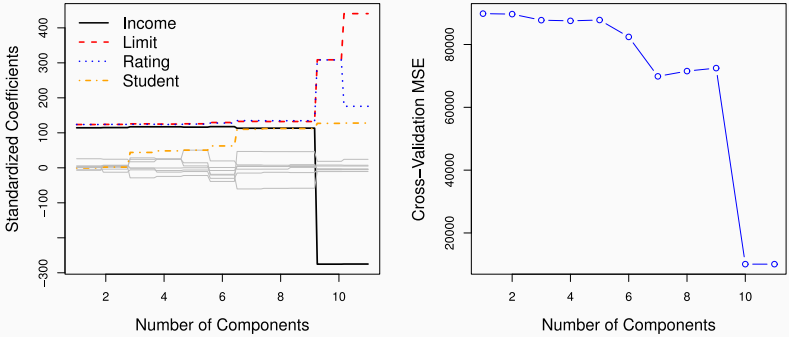
Grafy hodnot druhé hlavní komponenty z_{i2} *versus* pop *a* ad. *Vzájemné vztahy jsou slabé.*

Aplikace na regresi hlavních komponent



Regrese hlavních komponent (PCR - Principal Component Regression) byla použita na dva simulované soubory dat. Černá, zelená a purpurová čára odpovídají po řadě kvadrátu zkreslení, rozptylu a testovací střední kvadratické chybě. Vlevo: Simulovaná data ze slajdu 32. Vpravo: Simulovaná data ze slajdu 39.

Volba počtu směrů M



Vlevo: Normalizované odhady koeficientů PCR pro soubor dat `Credit` *pro různé hodnoty* M. Vpravo: MSE desetinásobné křížové validace získaná pomocí PCR jako funkce M.

Nevýhoda PCR

PCR určuje lineární kombinace, nebo *směry*, které nejlépe reprezentují prediktory $X_1, \ldots, X_p$.

- Tyto směry se určují způsobem bez učitele (nesupervizovaným), neboť odpověď Y se při stanovení směrů hlavních komponent nevyužívá.
- To jest, tato odpověď nesupervizuje identifikaci hlavních komponent.
- V důsledku toho PCR trpí potenciálně vážným nedostatkem: není zde žádná záruka, že směry, které nejlépe vysvětlují prediktory, budou také těmi nejlepšími směry, které by se měly použít k předpovídání odpovědi.

Částečné nejmenší čtverce

Částečné nejmenší čtverce (PLS, angl. *Partial Least Squares*) také stanovují nový soubor vlastností $Z_1, \ldots, Z_M$, které jsou lineárními kombinacemi původních vlastností, a pak těmito M novými vlastnostmi proloží lineární model pomocí obvyklých nejmenších čtverců.

- Na rozdíl od PCR identifikuje PLS tyto nové vlastnosti supervizovaně – využívá odpověď Y k nalezení nových vlastností, které *nejen dobře aproximují* staré vlastnosti, ale také *mají vztah k odpovědi.*
- Zhruba řečeno, přístup PLS usiluje o nalezení směrů, které pomáhají vysvětlit jak odpověď, tak prediktory.

Shrnutí

- Metody pro výběr modelu jsou podstatným nástrojem pro analýzu dat, obzvláště pro velké datové soubory zahrnující mnoho prediktorů.
- Výběr regresorů lze provádět s pomocí odhadů testovací chyby modelu, jako je C_p, AIC, BIC či upravené R^2, spolehlivější přístup spočívá ale v *křížové validaci*.
- Výzkum v oblasti metod, které dávají *řídkost*, jako je například *Lasso*, je obzvláště aktuální oblast. I zde se model typicky nastavuje pomocí křížové validace.
- Příbuzné přístupy, jako je například *elastická síť*, jsou popsány v knize.