11MAMY – Cvičení 8

Lineární regrese

Jan Přikryl

ČVUT FD

15. března 2023

Obsah cvičení

Teorie

Jednoduchá regrese na datech o automobilech

Vícenásobná regrese na datech o automobilech

Model s více regresory: prodeje dětských autosedaček

Jednoduché lineární regresní modely na syntetických datech

Kriminalita v Bostonu

Problém 1

Pečlivě vysvětlete rozdíly mezi KNN klasifikátory a KNN regresními metodami.

Problém 2

Shromáždíme sadu dat ($n = 100$ pozorování) obsahujících jediný prediktor a kvantitativní odpověď. Poté na datech identifikujeme lineární regresní model a také ještě kubickou regresi, tj. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$.

(a) Předpokládejme, že skutečný vztah mezi x a y je lineární, tj. $y = \beta_0 + \beta_1 x + \epsilon$. Zvažte trénovací zbytkový součet čtverců (RSS) pro lineární regresi a také trénovací RSS pro kubickou regresi. Budeme očekávat, že jedna hodnota bude nižší, než druhá, očekávali bychom, že budou stejné, nebo není dostatek informací k tomu, abychom to mohli říct? Odůvodněte odpověď.

(b) Odpovězte na (a) pro případ, kdy použijete RSS spočtené na testovací množině a nikoliv trénovací RSS.

Problém 2

pokračování

(c) Předpokládejme nyní, že skutečný vztah mezi x a y není lineární, že ale nevíme, jak daleko je od lineárního. Uvažujte trénovací RSS pro lineární regresi a také trénovací RSS pro kubickou regresi. Budeme očekávat, že jedna hodnota bude nižší, než druhá, očekávali bychom, že budou stejné, nebo není dostatek informací k tomu, abychom to mohli říct? Odůvodněte odpověď.

(d) Odpovězte (c) pro případ, kdy použijete RSS spočtené na testovací množině a nikoliv trénovací RSS.

Problém 3

Popište nulové hypotézy, kterým odpovídají p-hodnoty uvedené v tabulce. Vysvětlete, jaké závěry můžete vyvodit na základě těchto p-hodnot. Vaše vysvětlení by mělo být formulováno z hlediska `sales`, `TV`, `radio`, a `newspaper`, spíše než z hlediska koeficientů lineárního modelu.

	β_i	$s(\beta_i)$	t-statistika	p-hodnota
–	2,939	0,3119	9,42	$< 0{,}0001$
`TV`	0,046	0,0014	32,81	$< 0{,}0001$
`radio`	0,189	0,0086	21,89	$< 0{,}0001$
`newspaper`	-0,001	0,0059	-0,18	0,8599

Problém 4

Předpokládejme, že máme soubor dat s pěti prediktory, $x_1 =$ `GPA`, $x_2 =$ `IQ`, $x_3 =$ `Gender` (1 pro ženu a 0 pro muže), $x_4 =$ Interakce mezi `GPA` a `IQ` a $x_5 =$ Interakce mezi `GPA` a `Gender`. Závislou proměnnou je počáteční plat po promoci v tisících dolarů. Předpokládejme dále, že k sestavení modelu použijeme metodu nejmenších čtverců a dostaneme $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0{,}07$, $\beta_3 = 35$, $\beta_4 = 0{,}01$, $\beta_5 = -10$.

(a) Které z následujících tvrzení je správné a proč?

i. Pro danou hodnotu `IQ` a `GPA` vydělávají muži v průměru více, než ženy.
ii. Pro danou hodnotu `IQ` a `GPA` ženy vydělávají v průměru více, než muži.
iii. Pro danou hodnotu `IQ` a `GPA` vydělávají muži v průměru více, než ženy, pokud je `GPA` dostatečně vysoká.
iv. Pro danou hodnotu `IQ` a `GPA` vydělávají ženy v průměru více, než muži, za předpokladu, že je `GPA` dostatečně vysoká.

Problém 4

pokračování

(b) Predikujte plat ženy s IQ 110 a GPA 4,0.

(c) Pravda nebo nepravda:
Vzhledem k tomu, že koeficient pro interakci `GPA`/`IQ` je velmi malý, existuje velmi málo důkazů o interakčním účinku. Odůvodněte odpověď.

Automobily – jednoduchá regrese

Vyzoušejte si jednoduchou lineární regresi na datové sadě `islr_auto.csv`.

(a) Použijte funkci `mdl=fitlm()` pro stanovení jednoduché lineární regresní závislosti s `mpg` jako odezvou a `horsepower` jako prediktorem. Pro vypsání výsledků použijte obsah `mdl`. Komentujte výstup. Například:

i. Existuje nějaký vztah mezi prediktorem a odpovědí?
ii. Jak silný je vztah mezi prediktorem a odpovědí?
iii. Je vztah mezi prediktorem a odpovědí pozitivní nebo negativní?
iv. Jaké je předpokládané `mpg` spojené s `horsepower` 98? Jaké jsou 95 % intervaly spolehlivosti koeficientů a predikce?

Automobily – jednoduchá regrese

pokračování

Pokračujeme s grafy:

(b) Pomocí funkce `plot(mdl)` vykreslete odpověď a prediktor a zobrazte regresní přímku nejmenších čtverců.

(c) Pomocí funkce `coefCI(mdl)` vypište intervalové odhady jednotlivých parametrů regrese.

(d) Použijte funkci `mdl.plot()` pro vytvoření diagnostických grafů nejmenších čtverců. Komentujte jakýkoli problém s proložením dat přímkou, který zaznamenáte.

Automobily – vícenásobná regrese

Nyní se přesuneme k vícenásobné lineární regresi na téže datové sadě.

(a) Pomocí `gplotmatrix()` vytvořte matici korelačních diagramů, zahrnující všechny proměnné v datové sadě.

(b) Vypočtěte matici korelací mezi proměnnými pomocí funkce `corr()`. Budete muset vyloučit proměnnou `name`, která je kvalitativní?

(c) Použijte `mdl2=fitlm()` k určení vícenásobné lineární regrese s `mpg` jako odezvou a všemi ostatními proměnnými s výjimkou `name` jako prediktory. Pomocí `mdl2` vytiskněte výsledky. Komentujte výstup. Například:

i. Existuje vztah mezi prediktory a odpovědí?

ii. Které prediktory mají statisticky významný vztah k odpovědi?

iii. Co naznačuje koeficient pro proměnnou `year`?

Automobily – vícenásobná regrese

pokračování

Pokračujeme s grafy:

(d) Vytvořte diagnostické grafy lineární regrese. Komentujte jakýkoli problém, který vidíte s proložením. Naznačují grafy reziduí nějaké neobvykle velké odchylky? Vykazuje leverage graf nějaké pozorování s neobvykle vysokým pákovým efektem?

(e) Použijte symboly `"*"` a `":"` pro vytvoření lineárních regresních modelů s interakčními efekty. Jeví se nějaké interakce jako statisticky významné?

(f) Vyzkoušejte několik různých transformací proměnných, jako je například $\xi = \log(\mathbf{x})$, $\xi = \sqrt{\mathbf{x}}$, $\xi = \mathbf{x}^2$. Komentujte svá zjištění.

Carseats

Následující úkoly využívají simulovanou datovou sadu `islr_carseats.csv` o prodejích dětských autosedaček v supermarketech. Její atributy jsou

- `Sales`: počet prodaných kusů
- `CompPrice`: cena produktu od konkurence
- `Income`: místní úroveň příjmů (v tisících USD)
- `Advertising`: rozpočet na místní reklamu (v tisících USD)
- `Population`: počet obyvatel v regionu (v tisících)
- `Price`: prodejní cena za kus
- `ShelveLoc`: výhodnost umístění na regálu
- `Age`: průměrný věk místní populace
- `Education`: úroveň vzdělání v místě
- `Urban`: prodejna ve městě (ano/ne)
- `US`: sídlo obchodu je v USA (ano/ne)

Vyzkoušejte lineární regresi:

(a) Pro předpověď `Sales` pomocí `Price`, `Urban` a `USA` použijte model s více regresory.

(b) Interpretujte koeficienty v modelu. Buďte opatrní: některé proměnné v modelu jsou kvalitativní.

Carseats

pokračování

(c) Zapište model ve formě rovnice a dbejte na správné zacházení s kvalitativními proměnnými.

(d) Pro který z prediktorů lze odmítnout nulovou hypotézu $H_0 : \beta_j = 0$?

(e) Na základě vaší odpovědi na předchozí otázku použijte menší model, který používá pouze prediktory, u nichž existuje důkaz o jejich vlivu na výsledek.

(f) Jak dobře fungují modely uvedené v písmenech (a) a (e)?

(g) Pro model z (e) určete 95 % intervaly spolehlivosti pro koeficient(y).

(h) Existují v modelu podle (e) důkazy o výskytu odlehlých hodnot nebo o hodnotách s vysokým pákovým efektem (angl. *leverage*)?

Problém 7

V tomto cvičení budete vytvářet simulovaná data a budete na nich určovat jednoduché lineární regresní modely. Ujistěte se, že před začátkem části (a) použijete `rng(1)`, abyste zajistili konzistentní výsledky.

(a) Pomocí funkce `randn()` vytvořte vektor `x`, který obsahuje 100 pozorování z rozdělení $\mathcal{N}(0,1)$. Ten představuje příznaky X.

(b) Pomocí funkce `randn()` vytvořte vektor `eps`, který obsahuje 100 pozorování navzorkovaných z distribuce $\mathcal{N}(0,0{,}25)$, tj. z normálního rozdělení se střední hodnotou nula a rozptylem 0,25. Ten představuje šum ϵ.

(c) Pomocí `x` a `eps` vytvořte vektor `y` podle modelu

$$Y = -1 + 0{,}5X + \epsilon. \tag{1}$$

Jaká je délka vektoru `y`? Jaké jsou hodnoty β_0 a β_1 v tomto lineárním modelu?

Problém 7

pokračování

(d) Vytvořte scatterplot zobrazující vztah mezi `x` a `y`. Komentujte, co pozorujete.

(e) Pomocí metody nejmenších čtverců nalezněte model předpovídající `y` na základě `x`. Komentujte tento model. Jak se $\hat{\beta}_0$ a $\hat{\beta}_1$ shodují s β_0 a β_1?

(f) Do obrázku z bodu (e) zakreslete regresní přímku. Jinou barvou a tlustě vyznačte regresní přímku původní populace. Pomocí příkazu `legend()` vytvořte příslušnou legendu.

(g) Nyní vyzkoušejte polynomiální regresní model, který předpovídá `y` pomocí `x` a `x^2`. Existují nějaké důkazy, že by kvadratický člen zlepšil chování modelu? Vysvětlete svoji odpověď.

Problém 7

pokračování

(h) Opakujte (a)–(g) po úpravě procesu generování dat takovým způsobem, že je v těchto datech méně šumu. Model dle rovnice (1) by měl zůstat stejný. Lze toho dosáhnout snížením rozptylu normálního rozdělení použitého pro generování chybového členu `eps` v (b). Popište své výsledky.

(i) Opakujte (a)–(g) po úpravě procesu generování dat tak, aby data byla více zašumněná. Model dle rovnice (1) by měl zůstat stejný. Lze toho dosáhnout zvýšením rozptylu normálního rozdělení použitého pro generování chybového členu `eps` v (b). Popište své výsledky.

(j) Jaké jsou intervaly spolehlivosti pro β_0 a β_1 určené z původního souboru dat, více zašuměného souboru dat a méně zašuměné sady dat? Komentujte své výsledky.

Problém 8 – Kriminalita v Bostonu

Tento problém zahrnuje opět datovou sadu Boston, uloženou v `islr_boston.csv`, s níž jsme se již na cvičeních setkali. Pokusíme se na ní předpovědět míru kriminality na obyvatele za použití dalších proměnných v tomto souboru údajů. Jinak řečeno: míra kriminality na obyvatele je odpovědí a ostatní proměnné jsou předpovědi.

(a) Pro každý prediktor použijte jednoduchý model lineární regrese, který předpovídá odpověď. Popište své výsledky. V kterém z modelů existuje statisticky významná souvislost mezi prediktorem a odpovědí? Vytvořte nějaké obrázky, které vaše tvrzení podpoří.

(b) Použijte model s vícenásobnou regresí pro předpověď pomocí všech prediktorů. Popište své výsledky. Pro které prediktory lze odmítnout nulovou hypotézu $H_0 : \beta_j = 0$?

Problém 8 – Kriminalita v Bostonu

pokračování

(c) Jak se vaše výsledky z (a) srovnávají s výsledky z podle (b)? Vytvořte graf zobrazující jednorozměrné regresní koeficienty z (a) na ose x a více regresních koeficientů z (b) na ose y. To znamená, že každý prediktor je zobrazen jako jediný bod v grafu. Jeho součinitel v jednoduchém modelu lineární regrese je zobrazen na ose x a jeho odhad koeficientu v modelu vícenásobné lineární regrese je zobrazen na ose y.

(d) Existuje důkaz polynomiální závislosti mezi některým z prediktorů a odpovědí?

Chcete-li odpovědět na tuto otázku, použijte pro každý prediktor x_j model

$$y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3 + \epsilon.$$

a vyhodnoťte jej.

Dodatky

Problém 9

Uvažme proložené funkční hodnoty, jež jsou výstupem lineární regresní funkce bez absolutního členu. V tomto případě má i-tá proložená hodnota tvar

$$\hat{y}_i = \hat{\beta}_1 x_i$$

kde

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} x_i y_i \right) / \left(\sum_{j=1}^{n} x_j^2 \right).$$

Ukažte, že můžeme psát

$$\hat{y}_i = \sum_{j=1}^{n} a_j y_j.$$

a určete a_j.

Poznámka: Tento výsledek lze interpretovat tak, že výstupy lineární regrese jsou lineárními kombinacemi hodnot změřené odezvy.

Problém 10 a 11

Užitím vztahu (3.4) v knize (výpočet $\hat{\beta}_0$ a $\hat{\beta}_1$) potvrďte, že v případě jednoduché lineární regrese $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ prochází linie určená metodou nejmenších čtverců vždy bodem $(\bar{x}, \bar{y})$.

V textu knihy je také uvedeno, že v případě jednoduché lineární regrese y na x se R^2 statistická hodnota (3.17) rovná čtverci korelace mezi x a y (3.18). Dokažte, že tomu tak je. Pro jednoduchost můžete předpokládat, že $\bar{x} = \bar{y} = 0$.

Problém 12

V tomto úkolu budeme zkoumat t-statistiku nulové hypotézy $H_0 : \beta = 0$ v jednoduché lineární regresi bez absolutního členu. Začneme generovat prediktor x a odpověď y následujícím způsobem.

```
>> rng(42)
>> x = randn(100,1)
>> y = 2*x + randn(100,1)
```

(a) Proveďte jednoduchou lineární regresi y na x bez absolutního členu. Uveďte odhad koeficientu β, standardní chybu tohoto odhadu a hodnotu t-statistiky a p-hodnotu spojenou s nulovou hypotézou $H_0 : \beta = 0$. Komentujte výsledky.

(Regresi bez absolutního členu získáte použitím `fitlm(data,'y~x','Intercept',0)`.)

Problém 12

pokračování

(b) Nyní to obrátíme: Proveďte jednoduchou lineární regresi x na y bez absolutního členu a uveďte odhad koeficientu, jeho standardní chybu a odpovídající hodnotu t-statistiky a p-hodnotu spojené s nulovou hypotézou $H_0 : \beta = 0$. Komentujte výsledky.

(c) Jaký je vztah mezi výsledky získanými v bodech (a) a (b))?

(d) Pro regresi Y na X bez absolutního členu, má pro hypotézu $H_0 : \beta = 0$ odpovídající t-statistika tvar $\hat{\beta}/\,\mathrm{SE}(\hat{\beta})$, kde $\hat{\beta}$ je dána vztahem (3.38) z knihy a kde

$$\mathrm{SE}(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2}{(n-1)\sum_{j=1}^{n} x_j^2}}$$

Ukažte výpočtem a potvrďte numericky v Matlabu, že t-statistiku můžeme zapsat jako

$$\frac{\sqrt{n-1}\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{j=1}^{n} x_j^2)(\sum_{j=1}^{n} x_j^2) - (\sum_{i=1}^{n} x_i y_i)^2}}$$

Problém 12

pokračování

(e) Užitím výsledků z (d) potvrďte, že hodnota t-statistiky pro regresi y na x je stejná, jako hodnota t-statistiky pro regresi x na y.

(f) V Matlabu ukažte, že pokud uvažujeme i absolutní člen, je hodnota t-statistiky pro $H_0 : \beta_1 = 0$ stejná jak pro regresi y na x, tak i pro regresi x na y.

Problém 13

Opět pracujeme s jednoduchou lineární regresi bez absolutního členu.

(a) Připomeňme, že odhad koeficientu β pro lineární regresi Y na X bez absolutního členu je uveden v (3.38). Za jakých okolností je odhad koeficientu pro regresi X na Y stejný, jako odhad koeficientu pro regresi Y na X?

(b) V Matlabu vytvořte příklad s $n = 100$ pozorováními, v němž je odhad koeficientu pro regresi X na Y odlišný od odhadu koeficientu pro regresi Y na X.

(c) V Matlabu vytvořte příklad s $n = 100$ pozorováními, v němž je odhad koeficientu pro regresi X na Y totožný s odhadem koeficientu pro regresi Y na X.

Problém 14 – Kolinearita

Tento úkol se zaměřuje na problém kolinearity.

(a) Proveďte v Matlabu následující příkazy:

```
>> rng(11);
>> x1 = rand(100,1);
>> x2 = 0.5 * x1 + randn(100,1)/10;
>> y = 2 + 2 * x1 + 0.3 * x2 + randn(100,1);
```

(b) Poslední řádek odpovídá vytvoření lineárního modelu, v němž je `y` funkcí `x1` a `x2`. Napište rovnici tohoto lineárního modelu a jeho regresní koeficienty.

(c) Jaká je korelace mezi `x1` a `x2`? Vytvořte scatterplot zobrazující vztah mezi proměnnými.

Problém 14 – Kolinearita

pokračování

(d) Na těchto datech stanovte pomocí metody nejmenších čtverců regresní závislost pro předpověď `y` pomocí `x1` a `x2`. Popište získané výsledky. Co jsou β_0, β_1 a β_2? Jak se tyto vztahují ke skutečnému β_0, β_1 a β_2? Můžete odmítnout nulovou hypotézu $H_0 : \beta_1 = 0$? A co nulová hypotéza $H_0 : \beta_2 = 0$?

(e) Nyní stanovte regresní závislost předpovídající `y` pouze pomocí `x1`. Komentujte své výsledky. Můžete odmítnout nulovou hypotézu $H_0 : \beta_1 = 0$?

(f) Nyní stanovte regresní závislost předpovídající `y` pouze na základě `x2`. Komentujte své výsledky. Můžete odmítnout nulovou hypotézu $H_0 : \beta_1 = 0$?

(g) Jsou výsledky v bodech (d)–(f) navzájem v rozporu? Vysvětlete.

Problém 14 – Kolinearita

pokračování

(h) Nyní předpokládejme, že obdržíme jedno další pozorování, jež bylo bohužel nepřesné.

```
>> x1(end+1) = 0.1;
>> x2(end+1) = 0.8;
>> y(end+1) = 6;
```

Znovu stanovte lineární modely z (d)–(f) pomocí těchto nových dat. Jaký vliv má toto nové pozorování na každý model? Je v každém modelu toto pozorování odlehlé? Jde o bod s vysokým pákovým efektem? O obojí? Vysvětlete své odpovědi.