# Cvičení 9

## Lineární regrese

Jan Přikryl

ČVUT FD

8. března 2018

# Problém 1

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.

| | $\beta_i$ | $s(\beta_i)$ | $t$-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 2

Pečlivě vysvětlete rozdíly mezi KNN klasifikátory a KNN regresními metodami.

# Problém 3

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 3

## 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.

# Problém 4

Shromaždíme 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 tréninovací 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 (a) použitím testovacího spíše než trénovacího RSS.

# Problém 4

## 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) použitím testovacího spíše než trénovacího RSS.

# Problém 5

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

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

Co je $a_j$?

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

# Problém 6 a 7

Užitím vztahu (3.4) potvrďte, že v případě jednoduché lineární regrese prochází linie určená metodou nejmenších čtverců vždy bodem $(\bar{x}, \bar{y})$.

V textu knihy je 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 $\bar{x} = \bar{y} = 0$.

# Automobily – jednoduchá regrese

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

(a) Použijte funkci `mdl=fitlm()` pro provedení jednoduché lineární regrese s `mpg` jako odezvou a `horsepower` jako prediktorem. Pro vypsání výsledků použijte funkci `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 funkce `plot...` pro vytvoření diagnostických grafů nejmenších čtverců. Komentujte jakýkoli problém, který vidíte s proložením dat.

# 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.