Aplikovaná statistika v R - cvičení 3

Filip Děchtěrenko

Matematicko-fyzikální fakulta

filip.dechterenko@gmail.com

5.8.2014

Lineární regrese

- Obecně se snažíme popsat data pomocí rovnice (i odpovídá pozorvání)

$$data_i = (model) + chyba_i$$

- V lineární regresi vysvětlujeme jednu proměnnou pomocí dalších

$$Y_i = (b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n) + \epsilon_i$$

- Lineární je proto, že všechny b_i jsou lineární (žádné nadruhou či jiné funkce), namísto X_i může být cokoli, tedy toto je taktéž linerární regrese

$$Y_i = (b_0 + b_1 log(X_1) + b_2 X_2^2) + \epsilon_i$$

- Hodnotám ϵ_i se říká residuum – určuje odchylku jednotlivých pozorování (lidí) od modelu
- Zobecněním regrese i na jiné typy proměnných dostaneme Zobecněný linear model (General Linear Model; GLM)

Základní varianta - máme jen jednu spojitou proměnnou X_1

- Rovnice je tvaru
$$Y_i = b_0 + b_1 X_1 + \epsilon_i$$
- Pearsonův korelační koeficient mám popisuje, jak moc dobrá to je přímka
- Koeficientu b_0 říkáme **intercept**, koeficientu b_1 říkáme **slope**
- Koeficienty hledáme metodou nejmenších čtverců - hledáme takovou přímku, která nám mimimalizuje odchylky od přímky.
- Obvykle nás zajímá významnost jednotlivých koeficientů (tedy prediktorů), intercept je většinou nezajímavý (není překvapivé, že bude odlišný od nuly)

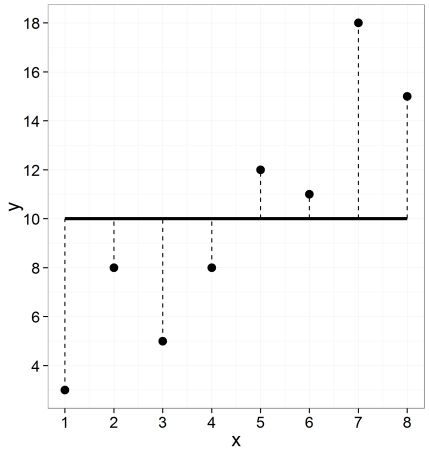
Goodness of fit – důležité odchylky

- **Goodness of fit** nám jedním číslem určuje, jak dobře náš model sedí, jde o vztah

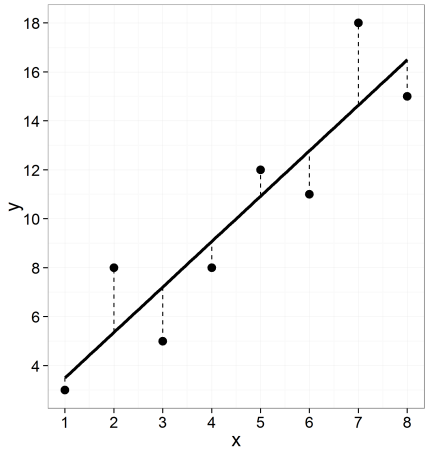
$$odchylka = \sum (pozorovani - model)^2$$

- Důležité odchylky
  - **Celková suma čtverců** (SS_T) – celkový součet odchylek od průměrné hodnoty Y
  - **Celková suma čtverců** (SS_R) – celkový součet reziduí
  - **Celková suma čtverců** (SS_M) – Rozdíl SS_T a SS_R, nám říká, o co je lepší naše predikce Y pomocí modelu než bez něj
- V případě regresní přímky nám to určuje parametr R^2, který známe z korelace. Spočítáme ho jako $R^2 = \frac{SS_M}{SS_T}$

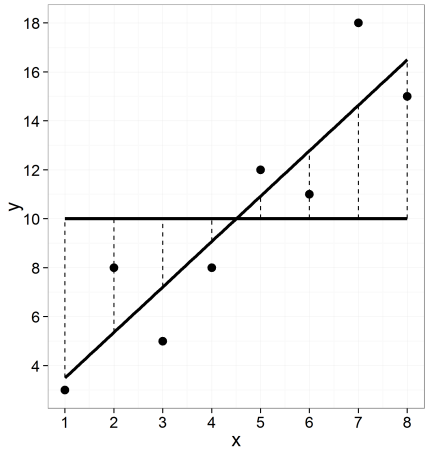
Goodness of fit – graficky



Obrázek : SS_T



Obrázek : SS_R



Obrázek : SS_M

Obrázek : Jednotlivé sumy čtverců

Vícerozměrná regrese

- Máme více proměnných X_i, předpokládáme, že všechny proměnné jsou spojité
- Takto nám může vzniknout několik modelů, v kterých se liší prediktory. Poté je můžeme porovnávat a určovat, zda má smysl přidávat další proměnné do modelu. Mluvíme o **hierarchické regresi**
- Mějme model $Plat = b_1 PocetOdpracHodin + \epsilon_i$ kde ϵ_i odpovídá osobnímu hodnocení. Má smysl do modelu přidat ještě pohlaví?
- Měli bychom mít 10-15 dat za každý prediktor (rule of thumb), ale záleží, jak velký efekt nás zajímá

Významná pozorování

- Některá pozorování ovlivňují směr přímky více než jiná. V datech se mohou nacházet odlehlá pozorování (**outliers**). Co s outliery je těžká otázka:-)
- Abychom určili, zda naše regrese je stabilní, a tedy zda dá i pro další modely stejné predikce, můžeme určit významná pozorování
- Je mnoho diagnostik, jak poznat významná pozorování, je dobré znát alespoň **Cookovu vzdálenost**, která by měla být menší než 1 (více prakticky)

Předpoklady regrese

- Prediktory musí mít nenulový rozptyl (nebývá problém)
- Prediktory spolu nesmí být velmi silně korelované (multicolinearity)
- Na různých úrovních prediktorů musí mít závislá proměnná stejný rozptyl (řeší se u GLM)
- Jednotlivé pozorování nesmí být spolu korelované (autocorrelation)
- (+několik triviálních podmínek)

Kategorický prediktor

- Regrese si snadno poradí i s prediktory, které nejsou spojité (třeba s pohlavím)
- Otázkou je, jakým způsobem je překódovat na čísla
- Nejjednodušším způsobem je *dummy coding*: máme-li n úrovní proměnné (třeba pohlaví má 2 úrovně), přidáme do modelu $n-1$ proměnných, kde každá z proměnných nám určuje, která kategorie je aktivní. Jedna z kategorií je braná jako baseline.
- Při testování významnosti jednotlivých prediktorů vždy testujeme významnost proměnné vůči referenční kategorii.

	X_1	X_2
Zelená	1	0
Červená	0	1
Bílá	0	0

Tabulka : Pro tři barvy zavedu dvě proměnné X_1 a X_2. Bílá je referenční úroveň

Konec cvičení