Úvod do lineární regrese

Supplementum ke cvičení 4ST201 Statistika

Lubomír Štěpánek[1, 2]

[1]Oddělení biomedicínské statistiky
Ústav biofyziky a informatiky
1. lékařská fakulta
Univerzita Karlova, Praha

ČVUT FBMI

[2]Katedra biomedicínské informatiky
Fakulta biomedicínského inženýrství
České vysoké učení technické v Praze

Obsah

Příklad

- V přiloženém souboru `_09_cviceni_.xlsx` jsou v záložce `mzdy_vs_vzdelani` vždy hodnoty mezd (v tisících korun) náhodně vybraných jedinců vzhledem k jejich dosaženému vzdělání. Existuje mezi dosaženým vzděláním a průměrnou výší mzdy na hladina významnosti $0{,}05$ závislost?

Příklad

- Ze 100 hodů jednou mincí padla hlava 60-krát. Je pravděpodobnost padnutí hlavy a orla shodná na hladině významnosti $0{,}05$?

Příklad

- U 6800 osob byla zjišťována barva očí a vlasů. Výsledky jsou uvedeny v následující tabulce.

		barva vlasů			
		světlá	kaštanová	černá	zrzavá
barva očí	světle modrá	1768	807	189	47
	šedá či zelená	946	1387	746	53
	tmavohnědá	115	438	288	16

Rozhodněme, zda barva očí a barva vlasů jsou navzájem závislé znaky.

Regresní analýza

- regresní analýza se zabývá jednostrannými závislostmi mezi jednou nebo více vysvětlujícími proměnnými $X_1, X_2, \ldots, X_p$ a vysvětlovanou proměnnou Y
- „jak X_1 ovlivňuje Y“, „jaký dopad má změna X_1 na hodnotu Y“
- závislost je popsána pomocí regresní funkce (přímka, parabola)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

kde koeficienty vyjadřují směr i „sílu“ závislosti a $\varepsilon \sim \mathcal{N}(0, 1^2)$ je chybová složka

- regresní funkci neznáme, odhadujeme ji za pomoci výběrového souboru

$$Y = b_0 + b_1 X$$

$$Y = b_0 + b_1 X + b_2 X^2$$

Odhad parametrů regresní funkce

- parametry regresní funkce jsou obvykle neznámé a je třeba je na základě hodnot výběrového souboru odhadnout
- volí se tzv. kritérium
- zpravidla pro odhad parametrů používáme metodu nejmenších čtverců, kdy minimalizujeme

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- existují i jiná kritéria, ale metoda nejmenších čtverců má výhodné statistické vlastnosti

Přímková regrese

- přímková regrese je popsána vztahem

$$Y = b_0 + b_1 X$$

  mezi závisle proměnnou Y a nezávisle proměnnou X
- odhady parametrů b_1 a b_0 jsou

$$b_1 = \hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{s_{xy}}{s_x^2}$$

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x}$$

- závislost obou proměnných měříme pomocí indexu korelace R či indexu determinace I

$$I^2 = R^2 = \frac{S_T}{S_Y}, \qquad I = \sqrt{I^2}$$

- posouzení kvality modelu lze testem o regresním parametru a testem o modelu

Test o regresním parametru

- testuje nulovou hypotézu $H_0 : \beta_j$ o tom, že j-tý lineární koeficient je roven nule, tedy že j-tý vysvětlující proměnná v modelu nehraje signifikantní roli
- testovým kritériem je

$$T = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} \sim t_{1-\alpha/2}(n-p)$$

kde $s_{\hat{\beta}_j}$ je směrodatná odchylka odhadu koeficientu $\hat{\beta}_j$ a platí

$$s_{\widehat{\beta}_j} = \sqrt{\frac{S_R}{n-p}} \cdot \sqrt{\frac{1}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}}$$

kde $S_R = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$ tak, že $\hat{y}_i$ je i-tá vyrovnaná hodnota, tedy hodnota proměnné Y odhadnutá regresní funkcí pro i-té pozorování

Test o regresním modelu

- testuje nulovou hypotézu

$$H_0 : \beta_0 = \text{konst.} \wedge \forall j \in \{1, 2, \ldots, p\} : \beta_j = 0$$

o tom, že všechny lineární koeficienty je rovny nule (a absolutní člen je konstantní), tedy že regresní model popisuje neexistující závislost mezi Y a $X_1, X_2, \ldots, X_p$

- testovým kritériem je

$$F = \frac{\frac{S_T}{p-1}}{\frac{S_R}{n-p}} \sim F_{1-\alpha}(p-1, n-p)$$

kde $S_T = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ je teoretický součet čtverců
a $S_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ je reziduální součet čtverců

Příklad

- V přiloženém souboru `_09_cviceni_.xlsx` jsou v záložce `udrzba_domu_vs_jeho_cena` vždy hodnoty nákladů na údržbu domu (v dolarech) a tržní cena domu (v tisících dolarů).
  - (i) Modelujme závislost nákladů na údržbu na ceně tržní domu regresní přímkou.
  - (ii) Zhodnoťme kvalitu modelu pomocí koeficientu determinace.
  - (iii) Interpretujme věcně hodnotu regresního koeficientu $\hat{\beta}_1$.
  - (iv) Odhadněme střední hodnotu nákladů u domů za 80 tisíc dolarů.

Literatura


Hindls, Richard, Stanislava Hronová, Jan Seger a Jakub Fischer. *Statistika pro ekonomy*. Praha: Professional Publishing, 2007. ISBN: 978-80-86946-43-6.


Marek, Luboš. *Statistika v příkladech*. Praha: Professional Publishing, 2015. ISBN: 978-80-7431-153-6.

Děkuji za pozornost!

lubomir.stepanek@vse.cz
lubomir.stepanek@lf1.cuni.cz
lubomir.stepanek@fbmi.cvut.cz

https://github.com/LStepanek/4ST201_Statistika