Aplikovaná statistika v R - cvičení 2

Filip Děchtěrenko

Matematicko-fyzikální fakulta

filip.dechterenko@gmail.com

5.6.2014

Přehled Rkových příkazů

Jen pro připomenutí nejzákladnější věci v Rku

- Pracujeme s data.frame, což je tabulka
- Data načítáme příkazem `read.csv`, jsou-li csv formátu s oddělovačem čárkou, jinak `read.csv2`
- Na prozkoumání data frame se hodí příkazy `dim`, `str` a `head`
- Na grafy `plot` a `hist`

Vztah dvou proměnných

- Chceme-li zkoumat, jak se dvě proměnné chovají společeně, můžeme to spočítat pomocí **kovariance**, v Rku příkaz `cov`
- Protože kovariance může nabývat různých hodnot, je vhodné ji normalizovat, čímž dostaneme **korelaci**
- Korelace nabývá hodnot mezi -1 a 1, přičemž 1 značí úplnou pozitivní korelaci, -1 úplnou negativní korelaci (jedna proměnná roste, druhá klesá)
- Máme 2 druhy koleací – dvě proměnné a parciální korelace
- **Bivariete correlations** je korelace mezi dvěma proměnnými, máme několik typů
  - Pearsonův korelační koeficient
  - Spearmanův korelační koeficient
  - Kendallův korelační koeficient
- **Parciální korelace** nám zachycuje vztah dvou proměnných kontrolujeme-li 3 proměnnou

Pearsonův korelační koeficient

- Předpoklady: intervalová proměnná, (normálně rozdělená data)
- Určuje míru lineární závislosti (tj. jak moc dobrá je přímka mezi daty)
- Můžeme testovat jeho významnost pomocí t-testu (Rko dělá za nás)
- Podle velikosti korelačního koeficientu můžeme mluvit o efektu (jde jen o doporučení, můžeme to interpretovat v rámci vlastního výzkumu)
  - $r = \pm.1$ Malý efekt
  - $r = \pm.3$ Střední efekt
  - $r = \pm.5$ Velký efekt
- **Koeficient determinace** R^2 určuje kolik procent variability jedné proměnné je sdíleno s druhou.

Odbočka – Testování hypotéz

- Ve vědě nemůžeme nic dokázat (kromě matematiky), můžeme pouze vyvracet!
- Když chceme něco otestovat, stanovíme si **nulovou**(H_0) a **alternativní**(H_A) hypotézu
- Vždy testujeme, jaká je pravděpodobnost, že naše naměřená data D pocházejí ze světa, kde platí nulová hypotéza H_0 ($P(D|H_0)$)
- Každý test má testovou statitiku, kterou porovnáváme s kritickou hodnotou (kdy by to bylo moc divné), resp. p-hodnotu s hladinou významnosti
- Nulová hypotéza je vždy ta situace, kdy se nic nděje. Jak by byla nulová hypotéza u testu významnosti korelačního koeficientu?

Chyba 1. a 2. druhu

- Při testování hypotéz mohou nastat čtyři případy

	H_0 platí	H_0 neplatí
Nezamítám H_0	OK	Chyba 2.druhu
Zamítám H_0	Chyba 1. druhu	OK

Tabulka : Chyba 1. a 2. druhu

- Snažíme se kontrolovat chybu 1. druhu, je to větší problém
- Pravděpodobnost chyby 2. druhu určuje **síla testu**

Ukázka reálného testu

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- P-hodnota je menší než 0.001, tedy zamítáme nulovou hypotézu na hladině významnosti 0.001 ve prospěch alternativní hypotézy (kterou tady neznáme)

Spearmanův korelační koeficient ρ

- Neparametrická statistika – pracuje s pořadím
- Hodí se, pokud data porušují normalitu a máme jich málo (< 20)
- Pokud se v proměnných vyskytuje moc stejných hodnot (ties), počítá jen odhad
- Má obdobně R_s^2

Kendallův korelační koeficient τ

- Neparametrická statistika
- Je lepší ho používat pro malé vzorky s množsvím shod
- Spearman je sice více populární, ale Kendall je lepším odhadem korelace v celé populaci
- Nemá koeficient determinace

Biserialní a bodově biseriální korelace

- Používají se v případě, že aspoň jedna proměnná je dichotomická (2 možnosti)
- **Bodově-biseriální** (r_{pb})korelace se používá, pokud jedna z proměnných je dichotomická a zároveň diskrétní (těhotoná/není těhotná)
- **Biseriální** (r_b) se liší od bodové tím, že jedna z proměnných je vlastně spojitá (prošel v testu/neprošel v testu)
- Bodově biseriální se matematicky rovná Pearsonově korelaci, takže prakticky to není třeba moc řešit
- Není dobré dichotomizovat data - vždy je lepší to počítat na spojité proměnné

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)

Konec cvičení