CVIČENÍ 4

13. březen 2017

Cvičení 1. Ze stránky `http://www.statsci.org/data/general/fullmoon.txt` získejte data `fullmoon`, zopakujte si, co znamenají jednotlivé proměnné a jaké jsou mezi nimi vztahy. Nafitujte v `R` model `model.0` ze cvičení z minulého týdne (model $\mathcal{F}_{H_0}$ z přednášky), t.j.

$$Y_i = \mu + \varepsilon_i, \quad i = 1, \ldots, N \tag{1}$$

kde $\varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$, a model `model.1` ze cvičení z minulého týdne (model $\mathcal{F}_{H_1}$ z přednášky), t.j.

$$Y_{ji} = \mu + \alpha_j + \varepsilon_{ji}, \quad j = 1, \ldots, J;\ i = 1, \ldots, n_j, \tag{2}$$

kde $\varepsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$.

Cvičení 2. S využitím teoretických výpočtů ze Cvičení 3 a 4 z minulého týdne spočtěte kvantity z těchto cvičení pro data `fullmoon` a najděte spočtené hodnoty ve výstupech z funkcí

(a) `summary(model.0)`;

(b) `summary(model.1)`;

(c) `anova(model.1)`;

(d) `anova(model.0, model.1)`;

(e) `aov`;

(f) `oneway.test`.

Cvičení 3. Uvažujte model $\mathcal{F}_{H_1}$ pro data `fullmoon`. Následující tabulka udává odhad koeficientu, příslušnou směrodatnou odchylku, pozorovanou hodnotu t statistiky pro test hypotézy o nulovosti koeficientu proti oboustranné alternativě a příslušnou p-hodnotu pro μ_{Before}, $\alpha_{\text{During}} - \alpha_{\text{Before}}$ a $\alpha_{\text{After}} - \alpha_{\text{Before}}$.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.9167     1.2138   8.994 2.15e-10 ***
MoonDuring    2.5000     1.7165   1.456    0.155
MoonAfter     0.5417     1.7165   0.316    0.754
```

Spočtěte stejnou tabulku pro μ_{Before}, μ_{During} a μ_{After}.

Cvičení 4. Uvažujte model $\mathcal{F}_{H_1}$ pro data `fullmoon`. Zopakujte si, jakou designovou matici `R` defaultně použije a ověřte si to pomocí příkazu `model.matrix`.

Použijete-li v zadání příkazu `lm`, části `formula` znak -1, zakážete `R` použít absolutní člen, t.j. zakážete, aby první sloupec designové matice byl sloupcem jedniček. S využití této volby můžete `R` přinutit, aby použilo designovou matici, která povede na parametrizaci modelu $\mathcal{F}_{H_1}$ přímo pomocí μ_{Before}, μ_{During} a μ_{After}, t.j.

$$Y_{ji} = \mu_j + \varepsilon_{ji}, \quad j = 1, \ldots, J;\ i = 1, \ldots, n_j. \tag{3}$$

Nafitujte tento model (nazvěme jej `model.2`) v `R` a porovnejte `summary(model.2)` s výsledky ze Cvičení 3 a se `summary(model.1)`.

Cvičení 5. Vraťme se nyní k definici modelu $\mathcal{F}_{H_1}$ pomocí (2), t.j. s koeficienty $\mu, \alpha_1, \ldots, \alpha_J$. Zopakujme si, že designová matice takového modelu má dimenzi $n \times (J+1)$ a hodnost J, a tedy existuje nekonečně mnoho voleb vektoru odhadů $(\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_J)$, které minimalizují součet čtverců $\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ji} - \hat{\mu} - \hat{\alpha}_j)^2$. Jednoznačně odhadnutelných parametrů v modelu $\mathcal{F}_{H_1}$ je J, konkrétně střední hodnoty jednotlivých skupin $\mu_1, \mu_2, \ldots, \mu_J$ a jejich lineární kombinace jsou jednoznačně odhadnutelné. Kdybychom ale model $\mathcal{F}_{H_1}$ chtěli parametrizovat přímo pomocí koeficientů $\mu_1, \ldots, \mu_J$, t.j. pomocí (3) dostali bychom model bez absolutního členu. Vzhledem k nevýhodám, které tato volba obnáší (viz Cvičení 4), obvykle volíme parametrizaci modelu $\mathcal{F}_{H_1}$ pomocí absolutního členu β_0 a dalších $J-1$ parametrů $\beta_1, \ldots, \beta_{J-1}$. Skupinové střední hodnoty $\mu_1, \mu_2, \ldots, \mu_J$ jsou funkcemi parametrů $\beta_0, \beta_1, \ldots, \beta_{J-1}$. Prozkoumejte vztah mezi $\beta_0, \beta_1, \ldots, \beta_{J-1}$ a $\mu_1, \mu_2, \ldots, \mu_J$. Uvědomte si, že vztah mezi $\beta_0, \beta_1, \ldots, \beta_{J-1}$ a $\mu_1, \mu_2, \ldots, \mu_J$ musí být vzájemně jednoznačný. V ANOVA se $\beta_1, \ldots, \beta_{J-1}$ často volí jako lineární kombinace $\mu_1, \mu_2, \ldots, \mu_J$, jejíchž koeficienty jsou kolmé na vektor jedniček (tvoří ortogonální kontrasty).

Cvičení 6. `R` umožňuje uživateli zvolit si parametrizaci $\beta_1, \ldots, \beta_{J-1}$. V parametru `contrasts` funkce `lm` můžeme přímo zadat, jak mají vypadat řádky designové matice pro pozorování z jednotlivých skupin. Jde vlastně o matici dimenze $J \times (J-1)$, kterou zprava pronásobíme designovou matici parametrizace (2) tak, abychom dostali novou designovou matici s plnou sloupcovou hodností[1]. Použijte znalosti získané ze Cvičení 5 na to, abyste si za pomoci volby `contrasts` nechali nafitovat model $\mathcal{F}_{H_1}$ parametrizovaný tak, aby

$$\begin{aligned} \beta_0 &= \mu, \\ \beta_1 &= \mu_{\text{During}} - \mu_{\text{Before}}, \\ \beta_2 &= \mu_{\text{After}} - \mu_{\text{During}}. \end{aligned}$$

Všimněte si, že koeficienty srovnávají po sebe jdoucí fáze měsíce. Zajímalo-li by nás jenom srovnání úplňku proti dvěma zbylým fázím, bylo by jednodušší použít volby `relevel` v části `formula`.

Domácí úloha (*15 bodů*)
Uvažujte model $\mathcal{F}_{H_1}$ pro data `fullmoon`, tentokrát ale modelujte vliv kalendářního měsíce na počet pacientů. Parametrizujte model tak, aby koeficienty odpovídaly srovnání zimních měsíců proti Vánocům a letních měsíců proti školním prázdninám. Přiložte zdrojový kód v `R` a výstup funkce `summary`.

Tip: User-friendly manuál k volbě kontrastů v `R` najdete například na této webové stránce: *`http://rstudio-pubs-static.s3.amazonaws.com/65059_586f394d8eb84f84b1baaf56ffb6b47f.html`.*

[1] Mluvili jsme o ní v přednášce z týdne 11 z podzimního semestru, částí o výběru řešení pro model s neúplnou hodností.