Předpoklady vybraných statistických testů

Obecně

- Většina předpokladů se týká:
  - **Dostatečné velikosti vzorku**.
  - **Nezávislosti reziduí**.
  - **Normálního rozdělení reziduí**.
  - **Linearity vztahů**.
  - **Homoskedasticity/shody rozptylů**.
  - **Absence extrémních/vlivných případů**.

Test nezávislosti chí-kvadrát

- **Pracujeme se dvěma kategorickými proměnnými**.
- **Obě z nich mají dvě nebo více úrovní**.
- **Pozorování/případy jsou nezávislé**. Typickým příkladem porušení tohoto předpokladu by bylo, kdybychom měli párová měření (např. pretest-posttest).
- **Dostatečná velikost vzorku**. Očekávaná četnost v každé buňce by měla být alespoň 1 a minimálně 80 % buněk by mělo mít očekávanou četnost 5 nebo více.
- V případě, že předpoklady týkající se minimálních očekávaných četností nebyly dodrženy, je možné použít:
    - **Fisherův exaktní test**. Vhodný u menších vzorků v řádu desítek osob. Je totiž výpočetně náročný, u velkých vzorků zabere výpočet dlouhou dobu.
    - **Simulaci Monte Carlo**. Vhodné u větších vzorků, kde by výpočet exaktního testu trval dlouhou dobu.

Nezávislý t-test

1. **Pracujeme se dvěma nezávislými skupinami** (typickým příkladem jsou muži a ženy).
2. **Závislá proměnná** by měla být měřena **minimálně na intervalové úrovni**.
3. V rámci obou skupin by měla závislá proměnná vykazovat přibližně **normální rozdělení**. Nároky na normalitu klesají s větší velikostí vzorku.
4. **Homogenita rozptylů**: závislá proměnná by měla v obou skupinách vykazovat podobný rozptyl. Welchův t-test tento předpoklad nevyžaduje.
5. **Absence extrémních případů**.

Párový (závislý) t-test

1. Pracujeme s **párem proměnných** (dvěma závislými měřeními), pro které má smysl počítat rozdíl (např. rozdíl mezi pretestem a posttestem, rozdíl mezi výškou bratra a sestry apod.).
2. Tyto proměnné jsou měřeny **minimálně na intervalové úrovni**.
3. **Rozdíly mezi oběma měřeními** (vypočteme-li novou proměnnou jako rozdíl mezi původními dvěma proměnnými) by měly mít **přibližně normální rozdělení**.
4. **V distribuci těchto rozdílů by se neměly vyskytovat extrémní případy**.

Lineární regrese

1. **Závislá proměnná je měřena minimálně na intervalové úrovni.**
2. **Počet případů musí být větší než počet prediktorů.** To je naprosté minimum. Doporučuje se mít několikanásobně více případů než prediktorů.
3. **Vztah mezi prediktory a závislou proměnnou je lineární**. Obvykle posuzujeme pomocí parciálních regresních grafů anebo matice bodových grafů (scatterplot matrix).
4. **Homoskedasticita**. Rozptyl reziduí by měl být konstantní. Obvykle ověřujeme pomocí bodového grafu predikovaných hodnota na ose X a reziduí na ose Y.
5. **Nezávislost reziduí**. Mají-li data hierarchickou strukturu, můžeme očekávat korelaci mezi rezidui. Ověřujeme pomocí Durbin-Watsonova testu (jeho testová statistika nabývá hodnot od 0 do 4, hodnoty okolo 2 indikují absenci autokorelace reziduí, hodnoty menší než 2 pozitivní autokorelaci, hodnoty vyšší než 2 negativní autokorelaci) nebo uložením reziduí a výpočtem samotné autokorelace.
6. **Normální rozdělení reziduí**. Ověřujeme obvykle pomocí histogramu nebo Q-Q či P-P grafu reziduí.
7. **Absence silné multikolinearity mezi prediktory**. Můžeme se podívat na korelace mezi prediktory a na statistiky VIF/Tolerance.
8. **Absence extrémních a vlivných případů**. Můžeme si jich všimnout v parciálních regresních grafech nebo grafu reziduí a predikovaných hodnot. Také si můžeme uložit různé statistiky extremity/vlivu, např. Cookovy vzdálenosti, a pomocí grafu se podívat, jestli některé případy výrazně "neodstakují" od ostatních (na arbitrání cut-off skóry bych se nespoláhal).

Binární logistická regrese

1. **Dichotomická (binární) závislá proměnná**. V SPSS zkontrolujte, že nabývá pouze hodnot 0/1, jinak SPSS převede všechny zbylé hodnoty (které nejsou 0 nebo 1) na 0.
2. **Nezávislost reziduí**. Rezidua si můžeme uložit a podívat se na jejich autokorelaci.
3. **Dostatečný počet případů** – pravidla palce:
   - N ≥ 10 × počet prediktorů / relativní četnost méně zastoupené úrovně závislé proměnné.
   - Navíc alespoň *n* = 5 v každé buňce při krostabulaci závislé proměnné a jednotlivých kategorických prediktorů. Při interakci dvou kategorických prediktorů (P1, P2) je nutné se podívat na četnosti v jednotlivých buňkách při krostabulaci P1 × P2 × Závislá proměnná.
4. **Absence silné kolinearity mezi prediktory**.
5. **Lineární vztah mezi spojitými prediktory a logitem závislé proměnné**.
6. **Absence vlivných případů**.

Mezisubjektová ANOVA/ANCOVA

1. **Závislá proměnná je měřena minimálně na intervalové úrovni**.
2. **Dostatečný počet případů**. Měli bychom zkontrolovat četnosti v jednotlivých buňkách při krostabulaci kategorických prediktorů (doporučuje se $n \geq 5$ v každé buňce).
3. **Nezávislost reziduí**.
4. **Absence silné multikolinearity mezi prediktory** (např. v důsledku nevyváženého designu). Při silnější multikolinearitě / nevyváženém designu může zvolený typ součtu čtverců zásadně ovlivnit odhad efektů jednotlivých prediktorů.
5. V rámci každé skupiny (v případě faktoriální ANOVA se tím myslí v rámci každé kombinace úrovní všech nezávislých proměnných) ověřujeme:
   A. **Normalitu rozdělení závislé proměnné**.
   B. **Absenci odlehlých případů**.
   C. **Shodu (homogenitu) rozptylů** (Welchova ANOVA tento předpoklad nevyžaduje).
6. **Klasická ANCOVA** navíc předpokládá:
   A. **Lineární vztah mezi spojitým prediktorem (kovariátem) a závislou proměnnou**.
   B. **Absenci interakce mezi kovariátem a kategorickými prediktory**.
   C. **Vyváženost skupin z hlediska úrovně kovariátu**.

Víceúrovňová lineární regrese I.

- **Platí všechny předpoklady "obyčejné" regrese**, ale ověřujeme několik dalších předpokladů.
- Nejprve je nutné uložit si rezidua úrovně 1. Použijeme k tomu argument /SAVE RESID pod příkazem MIXED.

1. **Rezidua úrovně 1 by měla mít přibližně normální rozdělení**. To lze ověřit pomocí histogramu či P-P/Q-Q grafu.
2. **Rezidua úrovně 1 by měla být vzájemně nezávislá.** Lze ověřit pomocí bodových grafů (vytvořených zvlášť pro každého hudebníka) s číslem měření na ose X a reziduem na ose Y. Shluky bodů by měly vypadat náhodně.
3. **Rezidua úrovně 1 by neměla záviset na hodnotách prediktorů úrovně 1.** To lze ověřit pomocí několika bodových nebo krabicových grafů s prediktory úrovně 1 na ose X a rezidui na ose Y.
4. **Rozptyl reziduí úrovně 1 by měl být v rámci každé jednotky úrovně 2 přibližně stejný**. To lze ověřit např. pomocí krabicových grafů s identifikátorem hudebníka na ose X a rezidui na ose Y.

Víceúrovňová lineární regrese II.

- Poté si uložíme rezidua 2 úrovně, což jsou vlastně odhady náhodných průsečíků a směrnic. Použijeme k tomu argument SOLUTION na konci řádku /RANDOM pod příkazem MIXED.

5. **Rezidua 2 úrovně by měla být vzájemně nezávislá**. Obtížně ověřitelné. Porušení tohoto předpokladu může nastat při opomenutí vyšší úrovně hierarchie dat (např. každý hudebník může "spadat" pod jiné instruktory, jiný hudební soubor atd.).

6. **Rezidua 2 úrovně by měla vykazovat multivariační normalitu**. Spokojíme se s ověřením univariační normality (ta je totiž podmínkou multivariační) pomocí histogramů nebo Q-Q/P-P grafu.

7. **Rezidua 2 úrovně by měla být nezávislá na hodnotách prediktorů úrovně 2**. To můžeme ověřit pomocí bodových nebo krabicových grafů (v závislosti na typu prediktoru).

8. **Rezidua úrovně 1 by neměla souviset s rezidui úrovně 2**. To můžeme ověřit pomocí bodových grafů (případně s loess křivkou).

9. **Prediktory úrovně 1 by měly být nezávislé na reziduích úrovně 2 a naopak prediktory úrovně 2 by měly být nezávislé na reziduích úrovně 1**. To lze ověřit pomocí krabicových nebo bodových grafů (v závislosti na typu prediktoru).