Binární logistická regrese

4. seminář

Binární logistická regrese

- Používá se, chceme-li predikovat, zda nastane určitý jev dichotomické povahy (např. vyhrál/prohrál, prospěl/neprospěl, přežil/nepřežil, uzdravil se/neuzdravil, onemocněl/neonemocněl atd.).
- Koeficienty logistické regrese *b* vyjadřují predikovanou změnu logaritmu šancí (logitu), že nastane sledovaný jev, při změně daného prediktoru o +1 úroveň/jednotu (za předpokladu, že hodnoty ostatních prediktorů zůstanou beze změny).
- Pro interpretaci velikosti efektu se častěji používá **poměr šancí (odds ratio, *OR*)**, který udává, o kolik se zvýší šance daného jevu při změně daného prediktoru o +1 jednotku (opět za předpokladu, že hodnoty ostatních prediktorů zůstanou beze změny).
  OR = (šance po zvýšení hodnoty prediktoru o +1 úroveň či jednotu) / (původní šance).
- OR = 1 značí nulový účinek (nikoli OR = 0), kdy se šance nemění.
- **V případě kvantitativních proměnných závisí OR samozřejmě na zvolené jednotce měření (např. sekundy/hodiny/dny/roky/století) – volte uvážlivě**.
- Pravidla palce: OR ≈ 1,68 (resp. 0,60) slabý; 3,47 (0,29) střední; 6,71(0,15) silný účinek.

Obecný postup zůstává stejný

1. Příprava, čištění a screening dat
2. Transformace, odvozené/vypočítané proměnné, rekódování
3. Popisné statistiky, vyjádření se k chybějícím datům
4. Plánované (konfirmační) analýzy
    a) ověření předpokladů
    b) testování plánovaných hypotéz / stanovení velikosti plánovaných efektů
5. Doplňkové, explorační analýzy

Dataset – obecné informace

- Budeme používat data, která obsahují informace a pasažérech lodi Titanic (ale ne o profesionální posádce).
- Primárním zdrojem pro vznik těchto dat byla **Encyclopedia Titanica**: https://www.encyclopedia-titanica.org/
- Data byla vytvořena pro účely soutěže, kde mají účastníci vytvořit co nejlepší model predikující přežití pasažéra Titanicu: https://www.kaggle.com/c/titanic

Dataset – přehled proměnných

Proměnná	Popis
PassengerId	ID pasažéra
Survived	Přežil pasažér nehodu Titatnicu?
Pclass	Jakou třídou pasažér cestoval (1., 2., nebo 3.; 1. třída byla samozřejmě ta "nejpřepychovější")
Name	Příjmení a jméno pasažéra
Sex	Pohlaví pasažéra
Age	Věk v letech
SibSp	Počet příbuzných z řad sourozenců a manželů/manželek na palubě
Parch	Počet dětí/rodičů na palubě
Ticket	Kód lístku
Fare	Poplatek za plavbu v librách
Cabin	Kód kajuty, kde pasažér přebýval
Embarked	Kód přístavu, kde se pasažér nalodil

Předpoklady, které je třeba zkontrolovat

- **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.
- **Nezávislost pozorování/reziduí** (uvažujte nad tím, proč zde bude zjevně narušena).
- **Dostatečný počet případů** – pravidla palce:
  - *N* alespoň = 5 × počet prediktorů / relativní četnost méně zastoupené úrovně závislé proměnné.
  - Navíc alespoň *n* = 5 v každé buňce při krostabulaci závislé proměnné a jednotlivých kategorických prediktorů.
- **Absence příliš silné kolinearity mezi prediktory**.
- **Lineární vztah mezi kvantitativními prediktory** (např. věk v našem modelu) **a logitem závislé proměnné**.
- **Absence vlivných případů**.

Tvorba modelu

- Použijeme data **titanic.sav**
- Budeme predikovat to, zda pasažér přežil, na základě věku, pohlaví a třídy, ve které pasažér cestoval.
- V **kroku 1** použijeme jako prediktory pouze pohlaví a třídu. V **kroku 2** přidáme jejich interakci. V **kroku 3** přidejte "exploračně", dle vlastního uvážení minimálně jeden další prediktor, který by podle Vás mohl mít efekt na pravděpodobnost přežití pasažéra Titanicu.
- **Na základě vlastního uvážení si stanovte specifické hypotézy týkající se věku, pohlaví, třídy a jejich interakce** – tj. zda očekáváte, že (1) větší pravděpodobnost přežití měli mladší, nebo starší pasažéři; (2) pasažéři "luxusnějších", nebo "ekonomičtějších" tříd; a (3) ve které třídě očekáváte největší a nejmenší rozdíl mezi muži a ženami v pravděpodobnosti přežití.

Ověření predikční schopnosti modelu

- **χ^2-testy ověřují nulovou hypotézu, že náš model nepredikuje lépe než model nulový (který by každému případu predikoval častěji zastoupenou kategorii závislé proměnné) nebo než model v předchozím kroku.**
- **Velikosti účinku pro model jako celek – pseudo-R^2. Kromě těch, která poskytuje SPSS, si můžeme spočítat také:**
- **McFaddenovo $R^2 = 1 - LL_{model}/LL_{null\_model}$, jehož logika vychází z toho, že LL (log-likelihood) je analogií reziduálního součtu čtverců z lineární regrese.**
- **Thujrovo R^2, které vypočteme jednoduše tak, že pro obě kategorie závislé proměnné vypočteme průměrnou predikovanou pravděpodobnost a poté rozdíl mezi oběma průměry. Logika je taková, že čím lépe náš model predikuje, tím větší by měly být predikované pravděpodobnosti pro případy s hodnotou = 1 závislé proměnné a tím menší by měly být predikované pravděpodobnosti pro případy s hodnotou = 0 závislé proměnné, a v důsledku toho by měl být rozdíl mezi průměry predikovaných pravděpodobností obou skupin co největší.**

Ověření predikční schopnosti modelu

- Dále můžeme posoudit:
  - **Přesnost** (accuracy): celkový podíl správně klasifikovaných případů.
  - **Senzitivitu** (recall): podíl případů správně predikovaných jako přeživších ze všech skutečných přeživších.
  - **Specificitu**: podíl případů správně predikovaných jako úmrtí ze všech skutečných úmrtí.
  - **Pozitivní prediktivní hodnotu** (precision): podíl správně predikovaných případů z těch případů, u kterých model predikoval přežití.
  - **Negativní prediktivní hodnotu**: podíl správně predikovaných případů z těch případů, u kterých model predikoval úmrtí.