PREDIKCE SRDEČNÍHO ONEMOCNĚNÍ POMOCÍ LOGISTICKÉ REGRESE

Srdeční onemocnění patří mezi nejčastější příčiny úmrtí na světě (Roth a kol., 2020). Mezi známé rizikové faktory patří vyšší věk, mužské pohlaví, kouření nebo vysoký cholesterol. V praxi se však tyto faktory často posuzují jednotlivě. Cílem této práce je zjistit, které z nich přispívají k predikci srdečního onemocnění, když jsou vzaty v úvahu všechny najednou. Konkrétně je testováno, zda věk a pohlaví zůstávají významnými prediktory i po kontrole klinických symptomů.

Data

Byl použit veřejně dostupný dataset Heart Disease z UCI (Janosi a kol., 1989), který obsahuje záznamy pacientů odeslaných na kardiologické vyšetření v Cleveland klinice. Po odstranění duplicitních záznamů obsahuje datová matice 302 pacientů ve věku 29 až 77 let (M = 54,4; SD = 9,0). V souboru bylo 206 mužů (68%) a 96 žen (32%). Srdeční onemocnění bylo diagnostikováno u 164 pacientů (54%), zbylých 138 (46%) bylo bez nálezu.

Metoda

K analýze byla použita logistická regrese. Cílem je predikovat binární výsledek - v tomto případě, zda pacient má nebo nemá srdeční onemocnění - na základě více prediktorů současně.

Výstupem logistické regrese je pro každý prediktor tzv. odds ratio (OR). Tato hodnota říká, jak se mění šance na onemocnění při změně daného faktoru o jednu jednotku:

- OR > 1 znamená, že faktor šanci na onemocnění **zvyšuje**.
- OR < 1 znamená, že faktor šanci **snižuje**.
- OR = 1 znamená, že faktor nemá žádný efekt.

Například OR = 6 znamená, že pacienti s daným faktorem mají 6× vyšší šanci na onemocnění. OR = 0,15 znamená, že šance je asi 7× nižší (1 / 0,15 ≈ 6,7).

Statistická významnost jednotlivých prediktorů byla posuzována Waldovým testem.

Hladina významnosti byla zvolena $\alpha = 0{,}05$.

Do modelu byly zařazeny následující proměnné:

Závislá proměnná:

- Přítomnost srdečního onemocnění (1 = ano, 0 = ne).

Regresory:

- **Věk** pacienta v letech.

Kovariáty:

- **Pohlaví** (muž/žena).
- **Hladina cholesterolu** v krvi (mg/dl). Vyšší hodnoty bývají považovány za rizikový faktor.
- **Maximální tepová frekvence** dosažená při zátěžovém testu (bpm). Vyšší hodnota naznačuje lepší kardiovaskulární kondici.
- **Typ bolesti na hrudi** - čtyři kategorie: asymptomatická (referenční), atypická angina, neanginální bolest a typická angina.
- **Přítomnost bolesti na hrudi při námaze** (ano/ne).

Výsledky

Výsledky logistické regrese shrnuje tabulka 1.

Tabulka 1: Výsledky logistické regrese

Prediktor	b	SE	z	p	OR
Věk	−0,038	0,020	−1,93	0,054	0,96
Pohlaví (muž vs. žena)	−1,891	0,380	−4,98	< 0,001	0,15
Cholesterol	−0,006	0,003	−2,00	0,046	0,99
Max. tepová frekvence	0,029	0,009	3,34	< 0,001	1,03
Bolest: atypická angina	1,868	0,485	3,85	< 0,001	6,48
Bolest: neanginální	1,826	0,383	4,77	< 0,001	6,21
Bolest: typická angina	1,665	0,562	2,97	0,003	5,29
Bolest při námaze (ano vs. ne)	−0,937	0,357	−2,63	0,009	0,39

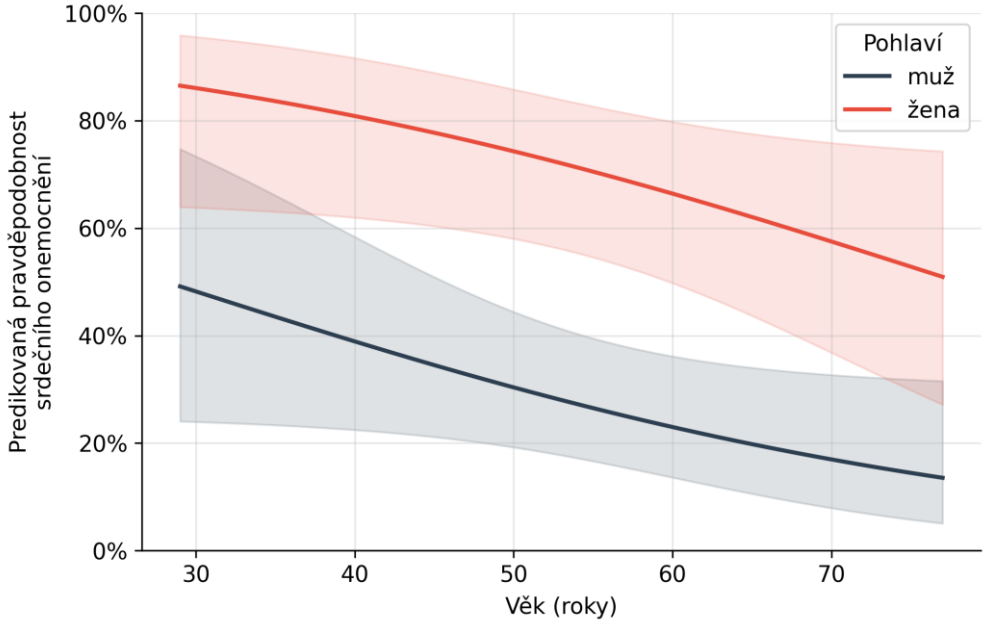
Nejsilnějším prediktorem je typ bolesti na hrudi. Pacienti s jakýmkoli typem bolesti mají přibližně 5-6× vyšší šanci na onemocnění oproti pacientům bez bolesti (asymptomatickým). Pohlaví je rovněž silně významné - muži mají v tomto vzorku výrazně nižší šanci na diagnózu (OR = 0,15), tedy přibližně 7× nižší než ženy. Vyšší maximální tepová frekvence mírně zvyšuje šanci na onemocnění (OR = 1,03 za každý bpm) a přítomnost bolesti při námaze ji naopak snižuje (OR = 0,39).

Věk se ukázal jako hraniční prediktor (p = 0,054) - těsně nesplňuje zvolenou hladinu významnosti. Cholesterol je těsně signifikantní (p = 0,046), ale jeho praktický efekt je zanedbatelný (OR = 0,99).

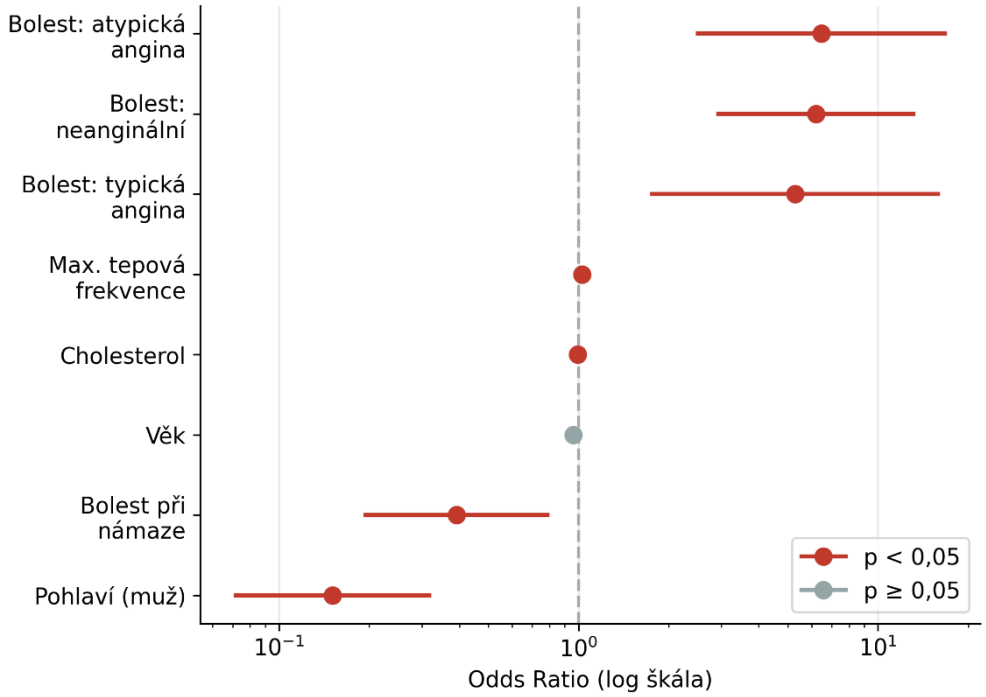
Model jako celek je statisticky významný (LR test: p < 0,001). Hodnota McFaddenova pseudo $R^2 = 0,36$ naznačuje přijatelnou schopnost modelu rozlišit mezi pacienty s onemocněním a bez něj.

Predikované pravděpodobnosti onemocnění v závislosti na věku a pohlaví zobrazuje obrázek 1. Přehled odds ratios s konfidenčními intervaly ukazuje obrázek 2.

Obrázek 1: Predikované pravděpodobnosti srdečního onemocnění podle věku a pohlaví



Obrázek 2: Forest plot - odds ratios jednotlivých prediktorů s 95% CI



Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=775

Interpretace a praktický význam

Výsledky ukazují, že přítomnost srdečního onemocnění je v tomto vzorku nejlépe predikována symptomy - především typem bolesti na hrudi. Naopak klasické rizikové faktory jako věk a cholesterol hrají po kontrole symptomů menší roli, než by bylo intuitivně očekáváno.

Překvapivý nález, že muži mají nižší šanci na diagnózu, lze vysvětlit povahou vzorku. Data nepocházejí z běžné populace, ale od pacientů, kteří již byli odesláni na kardiologické vyšetření. Muži bývají na kardiologii odesíláni častěji i preventivně, zatímco ženy spíše až při výraznějších příznacích. V důsledku toho ženy v tomto vzorku mají nižší zastopení, ale častěji skutečně potvrzené onemocnění.

Závěr

Logistická regrese se ukázala jako užitečný nástroj pro identifikaci faktorů spojených se srdečním onemocněním. Hlavním limitem práce je malý vzorek (N = 302) z jedné kliniky, což omezuje zobecnitelnost výsledků na širší populaci. Pro robustnější závěry by bylo vhodné analýzu replikovat na větším a populačně reprezentativnějším vzorku a na čerstvějších datech.

Literatura

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/

Roth, G. A., Mensah, G. A., Johnson, C. O., a kol. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019. *Journal of the American College of Cardiology, 76*(25), 2982–3021.