

Statistická analýza dat

disclaimer: žádná záruka správnosti -
nekontrolováno z žádným správným řešení

Jméno: _____

Podpis: _____

Cvičení	
Zkouška (písemná + ústní)	≥ 25
Celkem	≥ 50
Známka	

Pokyny k vypracování: doba řešení je 150min, jasně zodpovězte pokud možno všechny otázky ze zadání, pracujte s pojmy používanými v předmětu, můžete používat kalkulátory.

Statistické minimum. (10 b) Zodpovězte otázky níže. V každé z podotázk je právě jedna odpověď správná.

(a) (2 b) Pokud je p-hodnota dané testovací statistiky 0.03, pak můžeme:

- i) předpokládat, že extrémní hodnota testovací statistiky je způsobena náhodou,
- ii) dovedit, že za předpokladu platnosti nulové hypotézy existuje 3% pravděpodobnost získání extrémnější testové statistiky než jsme pozorovali,
- iii) předpokládat, že nulová hypotéza platí s pravděpodobností 0.03,
- iv) přijmout nulovou hypotézu na hladině významnosti 0.05.

(b) (2 b) Pokud chcete testovat hypotézu o střední hodnotě z populace s šíkým (nesymetrickým) rozdělením, měli byste:

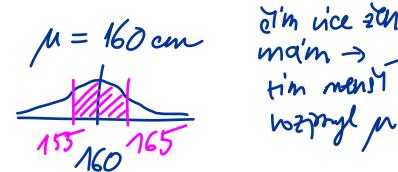
- i) použít stratifikovaný vzorek,
- ii) použít velkou hladinu významnosti α ,
- iii) mít všechny vzorky předem kategorizované jako úspěšné nebo neúspěšné,
- iv) získat vzorek s velikostí větší než 30.

(c) (2 b) Předpoklad nezávislosti při statistickém modelování znamená, že:

- i) nezávisle proměnné nejsou korelované,
- ii) chyby modelu spolu vzájemně nesouvisí,
- iii) nulová hypotéza by neměla být zamítnuta,
- iv) residua modelu nejsou nezávislá.

(d) (2 b) Víte, že výšky žen jsou normálně rozdeleny se střední hodnotou 160cm. Která z následujících pravděpodobností je nejvyšší? Pravděpodobnost náhodného výběru:

- i) jedně ženy s výškou mezi 155 a 165 centimetry,
- ii) 15ti žen s průměrnou výškou mezi 155 a 165 centimetry,
- iii) 100 žen s průměrnou výškou mezi 155 a 165 centimetry.
- iv) Tři výše uvedené jevy mají stejnou pravděpodobnost.



(e) (2 b) Chcete odhadnout podíl českých občanů, kteří podporují očkování proti Covidu. Jak velký vzorek (velikost značena n) by byl potřeba k zajištění 95% pravděpodobnosti, že skutečný podíl populace p nebude dále než 3 procentní body od podílu zjištěného ze vzorku? (tipy: počet příznivců očkování našem vzorku bude sledovat binomické rozdělení se střední hodnotou np a směrodatnou odchylkou $\sqrt{np(1-p)}$, měli byste pracovat s konzervativním odhadem, že $p = 0.5$, approximujte binomické rozdělení normálním, víte, že $z_{0,025} = 1.96$).

- i) 512,
- ii) 1068,
- iii) 2506,
- iv) 3152,
- v) 6304.

$$\text{binom. náhodný } k \text{ je } p \approx 95\%. \\ P(\text{true value} = \text{value} \pm 3\%) = 0,95 \\ p=0,5 \quad \text{zjištěná}$$

$$B_i \sim p(X=k) = \binom{n}{k} p^k \cdot (1-p)^{n-k} \\ E = np \\ \sigma = \sqrt{np(1-p)} \\ \text{počet příznivců } k \text{ a m polosedy}$$

$$1,96 \quad \sigma = 0,03 \\ z \quad \rightarrow \sigma = \frac{\text{směrodatná odchylka}}{m} = \sqrt{\frac{np(1-p)}{m}} = \sqrt{\frac{np(1-p)}{m^2}} = \sqrt{\frac{p(1-p)}{m}}$$

$$1,96 \quad \sqrt{\frac{p(1-p)}{m}} = 0,03$$

$$1,96^2 \frac{p(1-p)}{m} = 0,03^2 \rightarrow m = 1,96^2 \frac{p(1-p)}{0,03^2} = \underline{\underline{1064,1}}$$

Multivariátní regrese. (10 b) Sestavujete multivariátní lineární model. Nezávisle proměnných je velký počet, vyšší než je počet trénovacích vzorků, které máte k dispozici. Formálně tedy platí: $y \in R$, $\mathbf{x} \in R^p$, $T = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^m$, $p > m$, předpokládáme $y = \mathbf{x}^T \beta + \epsilon$.

(a) (2 b) Odhadněte a komentujte výsledek přímé aplikace metody nejmenších čtverců na množinu T .

moc prediktori - málo dat (m)
 $P < m$
 → nelze fungovat dobrě

satuovaný model - dokonalé proložení -
 přefitovávat - velká variace

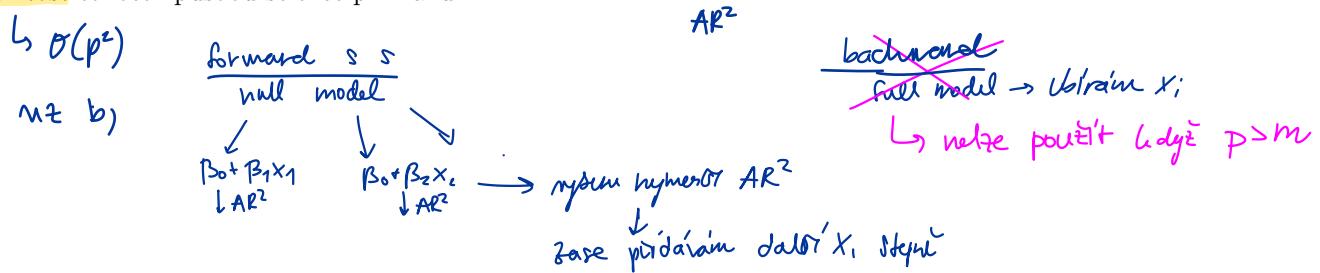
(b) (2 b) Popište co nejpodrobněji aplikaci metody výběru podmnožiny příznaků (subset selection) v souvislosti s tvorbou optimálního lineárního regresního modelu.

subset selection
 - snadno se rychlat odp model → získat subset příznaků
 - opakování aplikace LS na subseqy prediktori
 - musíme použít splávkové kritérium, které umí porovnat různé různé modely

(c) (2 b) Vysvětlete jaký problém ve srovnání s přístupem ad a) metoda ad b) řeší a zároveň vysvětlete, jaké úskalí naopak vzniká.

- řeší, že nemůže moc prediktori, které nic nesdílají ← povídáte regresy pro všechny modely a výsledky jsou horší, $O(2^P)$
 - různá výhoda že to není upřesně technicky daleko různé
 ↳ nelze overfittovat

(d) (2 b) Popište přístup k selekcii příznaků a regresi založený na dopředné a zpětné krokové selekcii příznaků (forward and backward stepwise regression). Na základě jakého kritéria modely srovnáváme? Odhadněte složitost tohoto způsobu selekce příznaků.



(e) (2 b) Který z dvojice přístupů ad d) byste upřednostnili v této konkrétní úloze a proč? Srovnejte také oba přístupy s metodou subset selection.

forward, protože backwards nelze použít pro $p > m$

Logistická regrese. (10 b) Diskutujte logistickou regresi.

(a) (2 b) Popište za jakých podmínek je využití logistické regrese vhodné (definujte úlohu včetně jejích variant daných různými typy proměnných).

logistická regrese je vhodná pro kategorické proměnné (**nominální**) - (barva očí) - {modrá, zelená, hnědá}

- zejména pro kategorický výstup y - dostanu pst mezi (0,1) --> převedení na klasifikaci
- nezávislá proměnná x může být jak spojitá, tak kategorická
- kromě kategorických proměnných, také **ordinální** (Po, Út, St...)

úloha multivariatních vztahů s kategorickými proměnnými - klasifikace do tříd, porozumění proměnným (proč jsme je zařadili do tříd)

$$C = \{c_1, c_2, \dots, c_n\} ; f: X \rightarrow C, f_p: X \times C \rightarrow \langle 0, 1 \rangle$$

(b) (3 b) Zapište definiční vztah logistické regrese. Vysvětlete význam proměnných. Srovnajte s lineární regresí (je logistická regrese **lineární** či **nelineární**, když selhává jedna metoda a když naopak druhá).

$$\text{logistická sigmoida} \quad P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

$$\text{logistická sigmoida} \quad P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

zobecněná lin. regrese - zahrnuje se jejich negativní vlastnosti

logistická regrese je lineární protože její dělící hranice je stále **lineární** ačkoliv je mapována pomocí nelin. funkce

lineární regrese je hodně citlivá na outliersy, na rozdíl od log. reg. *Při rozdílu na účelu*
nevýhodou log. reg. je horší **interpretabilita** koeficientů, oproti lin. reg.

(c) (2 b) Jaká je interpretace koeficientů v logistickém modelu (srovnajte s lineární regresí, kde koeficient vyjadřuje průměrnou změnu výstupu při jednotkové změně dané závisle proměnné a zafixování hodnot ostatních závisle proměnných)? *logit(p)*

interpretace je přímo těžko interpretovatelný, ale defacto je to nějaký posun, který je nutné udělat
horší interpretace, protože nelze říct změnu jako u lin. reg. (viz zadání), ale nyní je změna svázána přes logit funkci

$$\logit(p) = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{P(X=1)}{P(X=0)}\right) \rightarrow \underline{\text{log odds}}$$

success

$$\logit(p) = 0.5 + 0.13 * \text{study_hours} + 0.97 * \text{female}$$

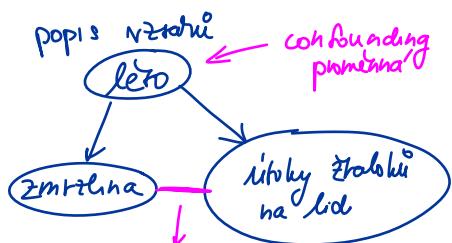
a 1 unit increase in X_1 will result in b increase in the log-odds of success : failure.

$$\begin{aligned} &X_1 + 1 \\ &\text{success} \rightarrow \exp(\beta_1) \\ &\text{failure} \times \exp(\beta_1) \end{aligned}$$

In the model above, $b = 0.13$, $c = 0.97$, and $p = P(Y=1)$ is the probability of passing a math exam. Let's pick study_hours and see how it impacts the chances of passing the exam. Increasing the study_hours by 1 unit (1 hour) will result in a 0.13 increase in $\logit(p)$ or $\log(p/(1-p))$. Now, if $\log(p/(1-p))$ increases by 0.13, that means that $p/(1-p)$ will increase by $\exp(0.13) = 1.14$. This is a 14% increase in the odds of passing the exam (assuming that the variable female remains fixed).

$$1.14 \text{ odds} = 14\% \text{ zvýšení}$$

(d) (3 b) Na konkrétním příkladu logistické regrese ilustrujte pojem matoucí proměnná (confounding variable).



která závisí a žuhok nemá korelace, ale je tomu korelace

Robustní statistika. (10 b) Předpokládejte, že sestavujete multivariátní lineární model $\hat{y} = \mathbf{x}^T \hat{\beta}$. Vaším cílem je odhadnout vektor parametrů modelu β z trénovacích dat.

(a) (2 b) Jak tento odhad provedete pokud předpokládáte, že skutečný vztah mezi závisle proměnnou Y a vektorem nezávisle proměnných \mathbf{X} lze popsat generativním modelem $Y = \mathbf{X}^T \beta + \epsilon$, kde ϵ je gaussovský šum $N(0, 1)$? Metodu pojmenujte a zapište její kritérium.

$$\text{MLE} \rightarrow \text{OLS} \quad \hat{\beta} = \arg \min$$

(b) (2 b) Zdůvodněte, proč je daná metoda v dané situaci optimální.

gaussovský šum

(c) (2 b) Jaké metody byste použili, pokud vztah zůstane nezměněn, ale šum ϵ bude směsí $\alpha N(0, 1) + (1 - \alpha)N(\gamma, 1)$, kde α je blízké 1 a γ je neznámé a konečné?

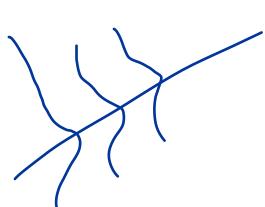
Hampel / Hader ?

(d) (2 b) Dojde vzhledem k předchozí situaci ke změně volby metody v případě, že ϵ je laplaceovský šum $\text{Laplace}(0, 1)$. Pokud ano, popište k jaké a proč.

→ median absolute regression

$$\sum |x_i^T \beta - y_i|$$

(e) (2 b) Proč nejsou metody popsané v bodě c) a d) vhodné i v situaci ad a)?



jiný šum → nelze dát přesný
ale medián

↳ min rozdíl w řadě
přesnou hodnotu

Shlukování. (10 b) Uvažujte problém shlukování.

- (a) (1 b) Definujte problém shlukování slovně.

vzdálení objektu x do k clusteru
s tím, že podobné objekty jsou uvnitř clusteru
odlišně jsou v jiném clusteru

- (b) (2 b) Formálně definujte shlukování jako optimalizační problém. Zařaďte jej do správné třídy složitosti. Zdůvodněte.

Bayesovská strategie $\arg \min_q \sum_{x \in X} p(x) \cdot W(x, q(x))$
→ procházení \forall strategií q → vybrat tě nejlepší strategii → NP hard protože
kutná phojit $\forall q$
→ proto se dělá heuristiky → strategie co se blíží k optimální

- (c) (3 b) Je spektrální shlukování příkladem shlukovacího algoritmu, který shlukování jako optimalizační problém formuluje a řeší? Vysvětlete.

ne! protože spektrální shlukování je aplikace nějakého běžného shlukovacího algoritmu na nový transformovaný příznakový prostor - zavádí nové transformace, ale mechanismus shlukování zůstává stejný (třeba k-means - a to je řešeno heuristicky)

- (d) (2 b) Používá spektrální shlukování kernel funkci? V čem se její použití liší od použití v kernel k-means?

ano, ale jinak!

kernel k-means je neefektivní kvůli výpočetní (časové) náročnosti - kvadratická složitost a

proto se moc nepoužívá

spektrální shlukování použije taky kernel, ale jako similarity funkci (matice), kterou pak dál upravuje pomocí Laplace a spektrálního rozkladu a až pak aplikuje k-means

- (e) (2 b) Popište, jak lze ve spektrálním shlukování nalézt optimální počet shluků k , pokud k není předem známo.

pomocí eigen-gap statistiky

hledám prudký skok ve velikosti vlastních čísel. Tolikáte vlastní číslo, kde je bod zlomu v hodnotách vlastních čísel, tam je optimální počet shluků k