

Zadání DÚ

Mgr. Zdeňka Geršlová

Příklad 1 - LRM (6 bodů)

V souboru `cneck.txt` máme k dispozici antropometrická data mladých dospělých lidí (převážně studentů VŠ z Brna a Ostravy). Chceme modelovat závislost tělesné hmotnosti (proměnná `body.W`, měřena v kg) na obvodu krku (proměnná `neck.C`, měřena v mm).

1. Načtěte data, prohlédněte si jejich strukturu a vypočítejte pro obě proměnné, se kterými budeme pracovat, výběrový průměr a směrodatnou odchylku.
2. Vykreslete bodový graf závislosti tělesné hmotnosti na obvodu krku (tj. graf, který má na ose x proměnnou `neck.C` a na ose y proměnnou `body.W`).
3. Sestavte lineární regresní model závislosti tělesné hmotnosti na obvodu krku. Pro vstup do modelu použijte datový soubor s odstraněným 36. pozorováním, které je odlehlou hodnotou. Prohlédněte si diagnostické grafy tohoto modelu a napište, zda jsou podle Vás splněny předpoklady pro lineární regresní model.
4. Vypište si podrobnosti o modelu a určete, zda je model jako celek statisticky významný a zda jako statisticky významný vychází intercept i obvod krku.
5. Vypočítejte intervaly spolehlivosti koeficientů modelu.
6. Vykreslete bodový graf (stejný jako v bodě 2) a proložte jej výslednou regresní přímkou.

Příklad 2 - PCA (6 bodů)

Pracujte s datovým souborem `du-kanga.txt`, který obsahuje údaje z měření lebek klokanů. Soubor obsahuje údaje o pohlaví (proměnná `sex`), druhu (proměnná `species`) a 12 rozměrů naměřených na lebkách. Vaším úkolem je provést analýzu hlavních komponent (PCA) pro **spojité** proměnné.

1. Načtěte datový soubor a prohlédněte si strukturu dat. Pokud jsou v souboru nějaké chybějící hodnoty, odstraňte příslušná pozorování. Vypočítejte korelační matici pro **spojité** proměnné.
2. Vypočítejte průměry pro všechny spojité proměnné a vykreslete přehled krabicových diagramů (tj. všechny diagramy v jednom obrázku). Rozhodněte na základě těchto údajů o tom, zda bude nutné v PCA použít škálování proměnných.
3. Proveďte PCA pro spojité proměnné (s nastavením škálování podle výsledků bodu 2). Vypište podíl variability a kumulativní podíl variability jednotlivých komponent a odpovězte na tyto otázky: Jaký podíl variability vysvětluje druhá hlavní komponenta? Jaký podíl variability je vysvětlen prvními třemi hlavními komponentami společně?
4. Rozhodněte o počtu hlavních komponent, se kterými budete nadále pracovat, podle Kaiserova kritéria. Kolik hlavních komponent bychom vybrali v případě, že bychom požadovali vysvětlení alespoň 80 % variability?
5. Vypočítejte korelaci původních proměnných s komponentami, které jste vybrali v předchozím bodě na základě Kaiserova kritéria.
6. Vykreslete pozorování a proměnné v rovině prvních dvou hlavních komponent (tzv. biplot). V grafu označte pozorování druhem příslušného jedince.