**Predikce spotřeby paliva**

**Milan Poláček**

# Zadání

Cílem tohoto úkolu je provést průzkumovou analýzu dat a modelovat závislost spotřeby auta na váze.

Data pro tento úkol byla modifikována z originálních dat v StatLib spravována Carnegie Mellon univerzitou. Originální data o autech byla nasbírána v roce 1980. U každého auta byla zjišťována, spotřeba, výkon, hmotnost, a další charakteristiky, viz Tabulka1

Data jsou k dispozici v souboru auta.csv

|  |  |  |
| --- | --- | --- |
| 1 | mpg | spotřeba - kolik mil lze ujet na galon paliva |
| 2 | cylinders | Počet válců |
| 3 | displacement | Velikost motoru |
| 4 | horsepower | Výkon |
| 5 | weight | Váha v librách |
| 6 | acceleration | Zrychlení |
| 7 | modelyear | Rok výroby |
| 8 | origin | Indikátor země původu |
| 9 | carname | Jméno auta |

Tabulka 1 Parametry přiložených dat

Požadované kroky analýzy:

1. Kolik máte k dispozici dat (kolik aut, jaké příznaky)? (0.5 bodu)

2. Obsahují data nějaké chybějící hodnoty? (0.5 bodu)

3. Jsou v datech nějaká odlehlá pozorování? Pokud ano, jak se s nimi vypořádáte?(1bod)

4. Vizualizujte vybrané příznaky, vztahy mezi příznaky vzhledem k ostatním bodům úkolu. (2 body)

5. Modelujte závislost spotřeby auta na váze auta a formálně ji zapište. Je statisticky významná? Výsledek slovně interpretujte. (4 body)

6. Existuje i závislost mezi spotřebou a jinými příznaky? (2 body)

# Řešení

K zadání úlohy byla přiložena data v nespecifickém formátu s různými odchylkami ve značení nenaměřených parametrů (příznaků).

Jak lze z tabulky 1 vyčíst máme 9 parametrů (příznaků), které byli měřeny u vozidel. Po úpravě dat do formátu CSV (Comma-separated values) a sjednocení značení chybějících pravděpodobně nenaměřených příznaků jsem zjistil tyto výsledky. Vozidel v poskytnutých datech je 398.

Jak jsem zmínil výše, v datech u některých vozidel chybí některé příznaky. Po úpravě značení chybějících příznaků jsem došel k tomu, že 5 vozidlům chybí data k velikosti motoru (displacement), 7 vozidlům chybí parametry o výkonu (horsepower) a 5 vozidlům chybí data o akceleraci (acceleration). Množiny dat těchto vozidel se nepřekrývají.

V datech lze nalézt měření, která jsou odlehlá, jako například dojezd vozidla na jeden galon paliva, kdy lze nalézt data se záporným dojezdem nebo absurdně vysokým dojezdem. Tyto data lze zanedbat v případě, že počet vyloučených vozidel ze statistiky je mnohem menší než celkový počet vozidel. U příznaku mpg se jedná o 4 vozidla (podle výše zmíněných kritérií), což je přibližně 1% z celkových dat a proto se dá předpokládat, že chyba nepřesáhne kritickou mez. Pro biologická data se považují testy s chybou 5% za dostatečně přesné.

|  |
| --- |
|  |
| Graf 1 Závislost spotřeby hmotnosti všech vozidel |

|  |
| --- |
|  |
| Graf 2 Závislost spotřeby hmotnosti vybraných vozidel |

Jak je vidět v grafu 1 jsou zde odlehlé hodnoty, jak jsem zmínil výše v příkladu o vyřazování dat s odlehlými příznaky mpg.

# Závěr

V této úloze jsme si potvrdili, že v modelování při zanedbání některých vlastností systému, můžeme získat zkreslené výsledky. Jako je například, že se nádoba zcela nevyprázdnila. Model zadaný v úloze je zcela funkční a bylo si na něm možné ověřit i základní fyzikální principy.

Řešení bonusové úlohy mi bohužel nevycházelo dle předpokládaných výsledků, a proto ho zde neuvádím.