# Generování datasetů pro různá statistická rozdělení

Vojtěch Müller
ČVUT - FIT
mullevo3@cvut.cz

17. května 2024

## 1 Úvod

Cílem je vytvořit Python nástroj pro generování datasetů dle různých statistických rozdělení. Součástí jsou také Jupyter notebooky ověřující správnost těchto generování.

## 2 Dataset

Datasetem se myslí matice o rozměrech $m \times n$, kde lineární kombinace jednotlivých sloupců a koeficientů tvoří vysvětlovanou proměnnou. Rozdělení vysvětlované proměnné se liší na základě zvolené distribuce.

Lze zvolit následující distribuce (parametr 'distribution'):

**norm** Gaussovo normální rozdělení

**poisson** Poissonovo rozdělení

**gamma** Gamma rozdělení

Dalšími konfigurovatelnými parametry jsou rozměry datasetu (*n_rows*, *n_columns*) a poměr informativních příznaků (tj. příznaků, ze kterých se skládá vysvětlovaná proměnná a tedy nejsou šumem) *informative_ratio*.

## 3 Proces generování

Generování probíhá stejně pro všechny distribuce a liší se až v posledním kroku.

1. **Vytvoření matice** – Úvodní vygenerování. Vygeneruje se matice o rozměrech $m \times n$, jehož prvky mají uniformní rozdělení $\sim$ Unif(-1, 1).
2. **Vytvoření interceptu** – Intercept je vybrán opět z rozdělení $\sim$ Unif(-1, 1).
3. **Výběr informativních příznaků** – Vyberou se náhodné příznaky, které jsou informativní (tj. ze kterých se pak bude počítat vysvětlovaná proměnná). Poměr informativních příznaků je parametrický.
4. **Vytvoření koeficientů** – Generování koeficientů z rozdělení $\sim$ Unif(-1, 1).
5. **Výpočet vysvětlované proměnné** – Vyberou se příznaky, které jsou informativní a pronásobí se s vygenerovanými koeficienty. Tato čísla se sečtou a vznikne vysvětlovaná proměnná *y_exact*.
6. **Transformace vysvětlované proměnné do vybraného rozdělení** – Dle vybrané distribuce se provede poslední krok. Proměnná *y_exact* se předá do *scipy* funkce *rvs* generující náhodné variety z vybraného rozdělení. Postupuje se podobně, jako v GLM modelu, tj. *y_exact* je prohnána skrz link function – u Poissona např. exponenciála.

## 4 Ověření vygenerování

V repozitáři se nacházejí jupyter notebooky ověřující vygenerované datasety pro každé rozdělení. K ověřování se použije vygenerovaný dataset, ze kterého se pomocí Bayesovského odhadu (knihovna *PyMC*) získají aproximace původních koeficientů použitých k vytvoření proměnné *y_exact*.

## 5 Limitace

Uvědomuji si, že v mém řešení je spousta limitací a beru jej spíše jako "proof of concept". Veškeré generování např. dělám z uniformního rozdělení z rozsahu $[-1, 1]$, což se rozhodně nepřibližuje variabilitě reálných datsetů. Toto řešení jsem použil zejména proto, protože např. u Poissonovské regrese se při generování dostávaly do exponenciály moc velká čísla a generování padlo. Toto by šlo vyřešit dělením *y_exact* nějakou konstantou $a$. Pokud bych pak pomocí GLM našel původní koeficienty a pronásobil je touto konstantou, dostal bych se k originálním.