Palmer Penguins: Statystyczna analiza danych

408453, Łukasz Wala, poniedziałek 14^{40}
AGH, Wydział Informatyki, Elektroniki i Telekomunikacji
Rachunek prawdopodobieństwa i statystyka 2021/2022

Kraków, 22 stycznia 2022

Ja, niżej podpisany własnoręcznym podpisem deklaruję, że przygotowałem przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.

.....

1 Streszczenie raportu

Raport powstał w oparciu o analizę cech 344 pingwinów zamieszkujących archipelag Palmera zebranych w latach 2007-2009.

2 Opis danych

Dane do projektu pochodzą z strony projektu `palmerpenguins`. Jest to projekt open-source mający dostarczyć dataset do wizualizacji i eksploracji danych będący alternatywą dla *Iris Dataset*, klasyka pośród materiałów do nauki statystycznej analizy danych i uczenia maszynowego.

W skład projektu wchodzą dwa zestawy danych: *penguins_raw* oraz jego uproszczona wersja *penguins*. W tym projekcie użyta zostanie wersja *penguins*. Dane zawierają 344 rekordy, gdzie każdy odpowiada innemu pingwinowi opisanemu ośmioma atrybutami:

- **species** - gatunek pingiwna (zmienna jakościowa spośród *Adelie*, *Chinstrap*, *Gentoo*),
- **island** - wyspa w archipelagu Palmera (zmienna jakościowa spośród *Biscoe*, *Dream*, *Torgersen*),
- **bill_length_mm** - długość dzioba (liczba zmiennoprzecinkowa, w milimetrach),
- **bill_depth_mm** - głębokość dzioba (liczba zmiennoprzecinkowa, w milimetrach),

- **flipper_length_mm** - długość płetwy (liczba całkowita, w milimetrach),
- **body_mass_g** - masa ciała (liczba całkowita, w gramach),
- **sex** - płeć (zmienna jakościowa spośród *male*, *female*),
- **year** - rok obserwacji (liczba całkowita)

Dane nie wymagają czyszczenia, ponieważ m.in. nie zawierają rekordów z nieznanymi wartościami czy odstającymi wartościami, wszystkie atrybuty będą istotne w analizie. Do załadowania ich do środowiska R można podejść na dwa sposoby. Z racji tego, że jest to zestaw danych przygotowanych z myślą o edukacji, został spakowany do pakietu R, który można zainstalować poprzez CRAN, a następnie załadować:

```
> install.packages("palmerpenguins", repos = "https://cloud.r-project.org/")
> library(palmerpenguins)
> data(package = 'palmerpenguins')
```

Wówczas do środowiska R dodany zostanie *dataframe* o nazwie *penguins*. Alternatywnym podejściem będzie pobranie pliku *.csv* z `repozytorium projektu` i załadowanie go:

```
> penguins <- read.csv(file = "penguins.csv", header = TRUE, )
```

3 Analiza danych

3.1 Wydobywanie podstawowych informacji z danych

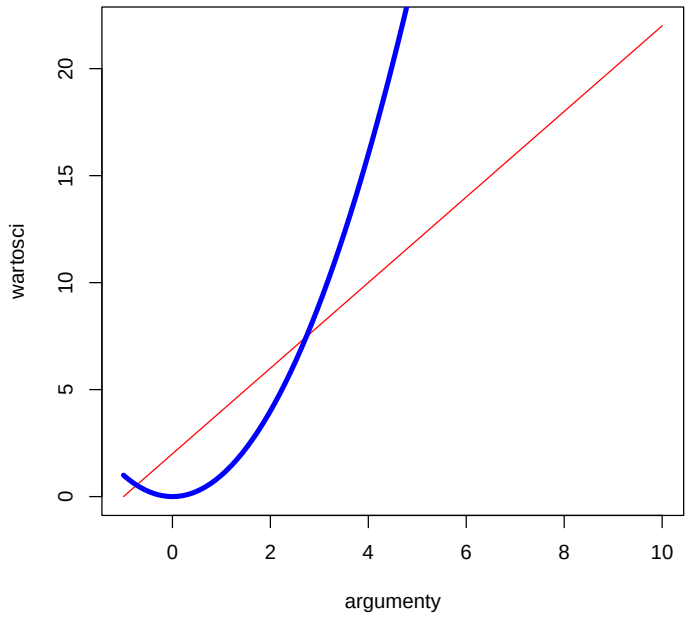
DziaĹ‚ania na liczbach, wartoĹ›ci funkcji w punkcie, zaokrÄ…glanie, dziaĹ‚ania logiczne.

```
> 5+7

[1] 12

> 3*4

[1] 12
```

3.2 Estymatory przedziaĹ‚owe

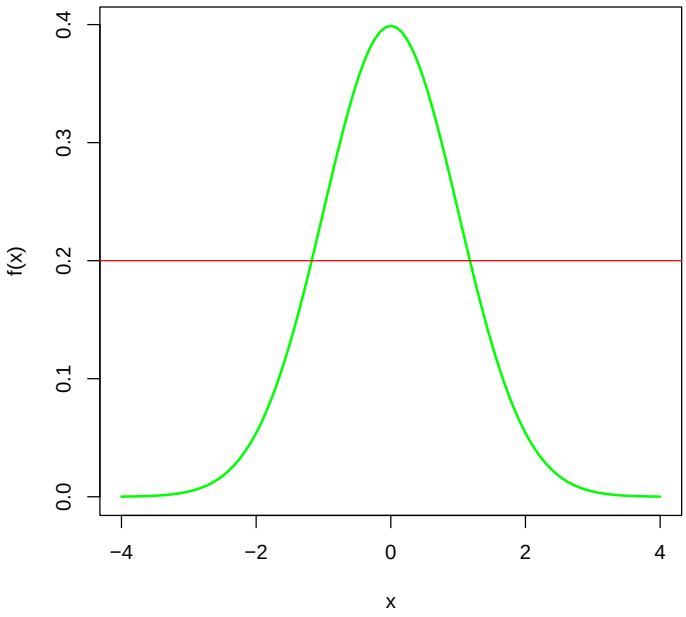
MoĹĽemy teĹĽ rysowaÄ‡, np. wykres funkcji $f(x) = 2x + 2$ oraz $g(x) = x^2$.

```
> f = function(x){2*x+2}
> g = function(x){x**2}
> curve(f, from=-1, to=10, xlab="argumenty", ylab="wartosci", col="red")
> curve(g, from=-1, to=10, xlab="x", ylab="y", col="blue", lwd=4, add=TRUE)
```



RĂłwnie dobrze moĹĽna narysowaÄ‡ gÄ™stoĹ■Ä‡ rozkĹ‚adu normalnego standardowego.

```
> curve(dnorm(x,0,1), from=-4, to=4, xlab="x", ylab="f(x)", col="green", lwd=2)
> abline(h=0.2,col="red")
```



3.3 Testowanie hipotez

TestowaÄ‡ bÄ™dziemy hipotezÄ™ zerowÄ… **H0**: ... wobec hipotezy alternatywnej **H1**: ... Korzystam ze statystyki t-Studenta postaci

$$t = \sum_{i=1}^{n} \frac{licznikX_i}{mianownik^2}$$

3.4 Regresja

W ten sposÄłb moĹĽna zapisaÄ‡ rÄłwnania w LaTeX, znakiem AND wyrÄłwnujemy je, a dwa slashe sĹ‚uĹĽÄ… do przejĹ■cia do kolejnej linii.

$$y = a \cdot x + b + \varepsilon,$$
$$z = 3 \cdot y.$$

4 Wnioski

Wnioski pĹ‚ynÄ…ce z przeprowadzonej analizy, sÄ… nastÄ™pujÄ…ce:

- wniosek pierwszy,

- wniosek drugi,
- i kolejne.