Statystyka wielowymiarowa
Laboratorium nr 1

Kamil Szkoła
Patryk Krukowski
Data Science, sem. 1

1 czerwca 2022

Spis treści

1 Wstęp

Naszym celem jest dokonanie analizy zbioru *Life.expectancy.csv* metodą regresji liniowej w celu przewidywania wartości *Life.expectancy*. Odpowiednie predyktory zostaną wybrane w dalszym ciągu analizy.

Sam zbiór danych został zebrany przez WHO w celu zbadania informacji dotyczących stanu zdrowia ludzi.

2 Analiza

2.1 Przygotowanie danych

Przed analizą musimy odpowiednio przygotwać dataset, ustawiając odpowiednie zmienne jako kategoryczne (funkcja *factor*).

```
> df = read.csv('life_expectancy.csv')
> df$Country <- factor(df$Country)
> df$Year <- factor(df$Year)
> df$Status <- factor(df$Status)
```

2.2 Ocena obecności liniowego wpływu predyktorów

Dopasujemy teraz model regresji liniowej do danych i ocenimy, czy którykolwiek z predyktorów ma liniowy wpływ na zmienną objaśnianą.

```
> fit_lm <- lm(Life.expectancy ~ . -Country -Year -Status, data = df)
> summary(fit_lm)

Call:
lm(formula = Life.expectancy ~ . - Country - Year - Status, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-17.0176  -2.0454  -0.0185   2.2260  11.9157

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.328e+01  7.358e-01  72.412  < 2e-16 ***
Adult.Mortality                 -1.689e-02  9.473e-04 -17.828  < 2e-16 ***
infant.deaths                    9.369e-02  1.068e-02   8.776  < 2e-16 ***
Alcohol                         -5.435e-02  3.061e-02  -1.776   0.0760 .
percentage.expenditure           3.777e-04  1.805e-04   2.093   0.0365 *
Hepatitis.B                     -5.582e-03  4.446e-03  -1.256   0.2095
Measles                         -8.617e-06  1.081e-05  -0.797   0.4253
BMI                              3.350e-02  6.011e-03   5.573 2.92e-08 ***
under.five.deaths               -7.047e-02  7.728e-03  -9.119  < 2e-16 ***
Polio                            7.836e-03  5.163e-03   1.518   0.1293
Total.expenditure                7.975e-02  4.074e-02   1.958   0.0505 .
Diphtheria                       1.439e-02  5.938e-03   2.423   0.0155 *
HIV.AIDS                        -4.383e-01  1.788e-02 -24.519  < 2e-16 ***
GDP                              1.383e-05  2.838e-05   0.487   0.6260
Population                      -6.917e-10  1.753e-09  -0.395   0.6931
thinness..1.19.years            -8.670e-03  5.310e-02  -0.163   0.8703
thinness.5.9.years              -5.123e-02  5.242e-02  -0.977   0.3286
Income.composition.of.resources  9.824e+00  8.340e-01  11.780  < 2e-16 ***
Schooling                        8.783e-01  5.939e-02  14.789  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.596 on 1630 degrees of freedom
  (1289 observations deleted due to missingness)
Multiple R-squared:  0.8347,        Adjusted R-squared:  0.8329
F-statistic: 457.4 on 18 and 1630 DF,  p-value: < 2.2e-16
```

Odrzucamy hipotezę H_0 na poziomie istotności 0.05, mówiącą o tym, że żaden z predyktorów nie ma liniowego wpływu na zmienną objaśnianą (*Life.expectancy*), bazując na F-statystyce.

2.3 Ocena istotności poszczególnych predyktorów

Z poprzedniego kodu R wynika, że na poziomie istoności równym 0.05 powinniśmy odrzucić następujące zmienne:

- Country,
- Year,
- Status,
- Alcohol,
- percentage.expenditure,
- Hepatitis.B,
- Measles,
- Polio,
- Total.expenditure,
- GDP,
- Population,
- thinness..1.19.years,
- thinness.5.9.years.

Nie mamy podstaw do odrzucenia pozostałych zmiennych. Dopasujemy teraz nowy model regresji liniowej, uwzględniający powyższe obserwacje.

```
> new_fit_lm <- lm(Life.expectancy ~.
+                  -Country
+                  -Year
+                  -Status
+                  - Alcohol
+                  - percentage.expenditure
+                  - Hepatitis.B
+                  - Measles
+                  - Polio
+                  - Total.expenditure
+                  - GDP
+                  - Population
+                  - thinness..1.19.years
+                  - thinness.5.9.years,
+                  data=df)
> summary(new_fit_lm)
```

```
Call:
lm(formula = Life.expectancy ~ . - Country - Year - Status -
    Alcohol - percentage.expenditure - Hepatitis.B - Measles -
    Polio - Total.expenditure - GDP - Population - thinness..1.19.years -
    thinness.5.9.years, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4652  -2.1147  -0.0277   2.1514  12.2507

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     52.4171654  0.5951021  88.081  < 2e-16 ***
Adult.Mortality                 -0.0178333  0.0009573 -18.628  < 2e-16 ***
infant.deaths                    0.0862870  0.0099406   8.680  < 2e-16 ***
BMI                              0.0377083  0.0057035   6.611 5.13e-11 ***
under.five.deaths               -0.0661585  0.0073981  -8.943  < 2e-16 ***
Diphtheria                       0.0139752  0.0046113   3.031  0.00248 **
HIV.AIDS                        -0.4327148  0.0180197 -24.013  < 2e-16 ***
Income.composition.of.resources 10.5985522  0.8281688  12.798  < 2e-16 ***
Schooling                        0.9531993  0.0558291  17.074  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.674 on 1640 degrees of freedom
  (1289 observations deleted due to missingness)
Multiple R-squared:  0.8264,        Adjusted R-squared:  0.8255
F-statistic: 975.8 on 8 and 1640 DF,  p-value: < 2.2e-16

>
```

Przejdziemy teraz do oceny charakterystyki wpływu każdego predyktora z osobna na odpowiedź modelu.

2.4 Ocena charakterystyki wpływu predyktorów

Charakterystyka wpływu na bazie poprzedniego kodu R:

- **Adult.Mortality** - ujemny, mały,
- **infant.deaths** - dodatni, mały,
- **BMI** - dodatni, mały,
- **under.five.deaths** - ujemny, mały,
- **Diphtheria** - dodatni, mały,
- **HIV.AIDS** - ujemny, mały,

- **Income.composition.of.resources** - dodatni, duży,
- **Schooling** - dodatni, mały.

Największy wpływ (dodatni) na zmienną objaśnianą ma zmienna **Income.composition.of.resources**.

2.5 Ocena dopasowania modelu do danych

Ocenę dopasowania modelu regresji liniowej ocenimy na podstawie miar R^2 oraz adjusted R^2, przy czym bardziej odpowiednią miarą jest adjusted R^2 z uwagi na uwzględnienie wzrostu liczby predyktorów w modelu (w przeciwieństwie do R^2).

```
> summary(new_fit_lm)$adj.r.squared

[1] 0.8255446

> summary(new_fit_lm)$r.squared

[1] 0.8263915
```

Obie miary wskazują na dobre dopasowanie modelu regresji liniowej do danych.