Metody przetwarzania i analizy danych w R

Łukasz Wawrowski

Contents

Wprowadzenie

Literatura podstawowa:

- Przemysław Biecek - *Przewodnik po pakiecie R*
- Marek Gągolewski - *Programowanie w języku R. Analiza danych, obliczenia, symulacje.*
- Garret Grolemund, Hadley Wickham - *R for Data Science* (polska wersja)

Literatura dodatkowa:

- inne pozycje po polsku
- inne pozycje po angielsku

Internet:

- R-bloggers
- rweekly

Chapter 1

Wprowadzenie

1.1 Narzędzie

- darmowe
- wszechstronne
- wsparcie społeczności
- wersja desktopowa i serwerowa

czyli **R** - środowisko do obliczeń statystycznych i wizualizacji wyników

- strona projektu: r-project.org
- świetne IDE: RStudio
- wersja przeglądarkowa: rstudio.cloud

R + Python

1.2 Cele analiz

Podstawowe:

- wnioskowanie statystyczne - porównywanie grup
- regresja - poszukiwanie związków
- klasyfikacja - przyporządkowanie do grup
- grupowanie - poszukiwanie grup
- prognozowanie - patrzenie w przyszłość

Inne:

- analiza języka naturalnego
- rozpoznawanie obrazów
- analiza koszykowa
- ...

1.2.1 Eksporacja danych

Pakiet `tidyverse`

```
library(tidyverse)
```

- analiza częstości dla zmiennych jakościowych
- analiza struktury dla zmiennych ilościowych

Case study: Wybory 2018

Chapter 2

Testowanie hipotez

2.1 Hipoteza statystyczna

Przypuszczenie dotyczące własności analizowanej cechy, np. średnia w populacji jest równa 10, rozkład cechy jest normalny.

Formułuje się zawsze dwie hipotezy: hipotezę zerową (H_0) i hipotezę alternatywną (H_1). Hipoteza zerowa jest hipotezą mówiącą o równości:

$H_0 : \bar{x} = 10$

Z kolei hipoteza alternatywna zakłada coś przeciwnego:

$H_1 : \bar{x} \neq 10$

Zamiast znaku nierówności ($\neq$) może się także pojawić znak mniejszości ($<$) lub większości ($>$).

2.2 Poziom istotności i wartość p

Hipotezy statystyczne weryfikuje się przy określonym poziomie istotności α, który wskazuje maksymalny poziom akceptowalnego błędu (najczęściej $\alpha = 0,05$).

Większość programów statystycznych podaje w wynikach testu wartość p. Jest to prawdopodobieństwo uzyskania obserwowanych wyników przy założeniu prawdziwości hipotezy zerowej.

Generalnie jeśli $p < \alpha$ - odrzucamy hipotezę zerową.

Krytyka wartości p

2.3 Testy parametryczne i nieparametryczne

Testy statystyczne dzielą się na dwie grupy:

- parametryczne, które wymagają spełnienia założeń, ale są dokładniejsze,
- nieparametryczne, które nie wymagają tylu założeń, ale są mniej dokładne.

Chapter 3

Regresja

3.1 Regresja prosta

Na podstawie danych dotyczących informacji o doświadczeniu i wynagrodzeniu pracowników zbuduj model określający 'widełki' dla potencjalnych pracowników o doświadczeniu równym 8, 10 i 11 lat.

regresja_prosta.Rmd

cały projekt

3.1.1 Zadanie

Dla danych dotyczących sklepu nr 77 opracuj model zależności sprzedaży od liczby klientów. Ile wynosi teoretyczna sprzedaż w dniach, w których liczba klientów będzie wynosiła 560, 740, 811 oraz 999 osób?

3.2 Regresja wieloraka

Na podstawie danych dotyczących zatrudnienia opracuj model, w którym zmienną zależną jest bieżące wynagrodzenie. Jaka cecha ma największy wpływ na tę wartość?

Opis zbioru:

- id - kod pracownika
- plec - płeć pracownika (0 - mężczyzna, 1 - kobieta)
- data_urodz - data urodzenia
- edukacja - wykształcenie (w latach nauki)
- kat_pracownika - grupa pracownicza (1 - ochroniarz, 2 - urzędnik, 3 - menedżer)
- bwynagrodzenie - bieżące wynagrodzenie
- pwynagrodzenie - początkowe wynagrodzenie
- staz - staż pracy (w miesiącach)
- doswiadczenie - poprzednie zatrudnienie (w miesiącach)
- zwiazki - przynależność do związków zawodowych (0 - nie, 1 - tak)
- wiek - wiek (w latach)

Wczytanie bibliotek `tidyverse`, `readxl` oraz danych.

```
library(tidyverse)
library(readxl)
```

```
options(scipen = 100)

pracownicy <- read_xlsx("data/pracownicy.xlsx")

pracownicy2 <- pracownicy %>%
  filter(!is.na(wiek)) %>%
  select(-id, -data_urodz) %>%
  mutate(plec=as.factor(plec),
         kat_pracownika=as.factor(kat_pracownika),
         zwiazki=as.factor(zwiazki))

summary(pracownicy2)
```

```
##  plec       edukacja     kat_pracownika bwynagrodzenie   pwynagrodzenie
##  0:257   Min.   : 8.00   1:362          Min.   : 15750   Min.   : 9000
##  1:216   1st Qu.:12.00   2: 27          1st Qu.: 24000   1st Qu.:12450
##          Median :12.00   3: 84          Median : 28800   Median :15000
##          Mean   :13.49                  Mean   : 34418   Mean   :17009
##          3rd Qu.:15.00                  3rd Qu.: 37050   3rd Qu.:17490
##          Max.   :21.00                  Max.   :135000   Max.   :79980
##       staz       doswiadczenie    zwiazki      wiek
##  Min.   :63.00   Min.   :  0.00   0:369   Min.   :24.00
##  1st Qu.:72.00   1st Qu.: 19.00   1:104   1st Qu.:30.00
##  Median :81.00   Median : 55.00           Median :33.00
##  Mean   :81.14   Mean   : 95.95           Mean   :38.67
##  3rd Qu.:90.00   3rd Qu.:139.00           3rd Qu.:47.00
##  Max.   :98.00   Max.   :476.00           Max.   :66.00
```

W zmiennej wiek występował brak danych, który został usunięty. Usunięto także kolumny, które nie przydadzą się w modelowaniu. Ponadto dokonujemy przekształcenia typu cech, które są jakościowe (płeć, kat_pracownika, zwiazki) z typu liczbowego na czynnik/faktor, który będzie poprawnie interpretowany przez model.

W modelu zmienna zależna to `bwynagrodzenie`, natomiast jako zmienne niezależne bierzemy pod uwagę wszystkie pozostałe cechy.

```
model <- lm(bwynagrodzenie ~ ., pracownicy2)
summary(model)
```

```
##
## Call:
## lm(formula = bwynagrodzenie ~ ., data = pracownicy2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23185  -3041   -705   2591  46295
##
## Coefficients:
##                    Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)     -4764.87418  3590.49652  -1.327              0.18514
## plec1           -1702.43743   796.51779  -2.137              0.03309 *
## edukacja          482.43603   160.83977   2.999              0.00285 **
## kat_pracownika2  6643.17910  1638.06138   4.056  0.00005869172850407 ***
## kat_pracownika3 11169.64519  1372.73990   8.137  0.00000000000000377 ***
## pwynagrodzenie      1.34021     0.07317  18.315 < 0.0000000000000002 ***
```

```
## staz              154.50876    31.65933   4.880  0.00000145958620443 ***
## doswiadczenie     -15.77375     5.78369  -2.727              0.00663 **
## zwiazki1        -1011.55276   787.80884  -1.284              0.19978
## wiek              -64.78787    47.88015  -1.353              0.17668
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6809 on 463 degrees of freedom
## Multiple R-squared:  0.8444, Adjusted R-squared:  0.8414
## F-statistic: 279.1 on 9 and 463 DF,  p-value: < 0.00000000000000022
```

Tak zbudowany model wyjaśnia 84% zmienności bieżącego wynagrodzenia, ale nie wszystkie zmienne są w tym modelu istotne.

Parametry regresji mają następujące interpretacje:

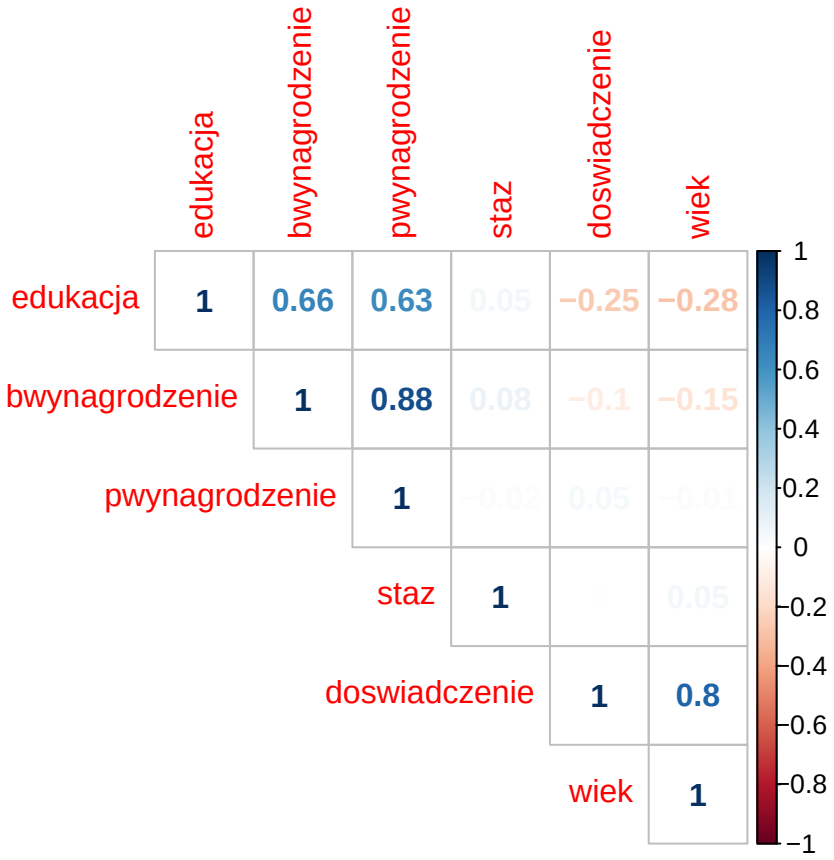
- plec1 - kobiety zarabiają przeciętnie o 1702,44 zł mniej niż mężczyźni,
- edukacja - wzrost liczby lat nauki o rok powoduje średni wzrost bieżącego wynagrodzenia o 482,44 zł
- kat_pracownika2 - pracownicy o kodzie 2 (urzędnik) zarabiają średnio o 6643,18 zł więcej niż pracownik o kodzie 1 (ochroniarz)
- kat_pracownika2 - pracownicy o kodzie 3 (menedżer) zarabiają średnio o 11169,65 zł więcej niż pracownik o kodzie 1 (ochroniarz)
- pwynagrodzenie - wzrost początkowego wynagrodzenia o 1 zł powoduje przecięny wzrost bieżącego wynagrodzenia o 1,34 zł
- staz - wzrost stażu pracy o miesiąc skutkuje przeciętnym wzrostem bieżącego wynagrodzenia o 154,51 zł
- doswiadcznie - wzrost doświadczenia o miesiąc powoduje średni spadek bieżącego wynagrodzenia o 15,77 zł
- zwiazki1 - pracownicy należący do związków zawodowych zarabiają średnio o 1011,55 zł mniej aniżeli pracownicy, którzy do związków nie zależą
- wiek - wzrost wieku pracownika o 1 rok to przecięnym spadek bieżącego wynagrodzenia o 64,79 zł

Wszystkie te zależności obowiązują przy założeniu *ceteris paribus* - przy pozostałych warunkach niezmienionych.

Ten model wymaga oczywiście ulepszenie do czego wykorzystamy m.in. pakiet olsrr.

Pierwszą kwestią, którą się zajmiemy jest współliniowość zmiennych. W regresji zmienne objaśniające powinny być jak najbardziej skorelowane ze zmienną objaśnianą, a możliwie nieskorelowane ze sobą. W związku z tym wybieramy ze zbioru wyłącznie cechy ilościowe, dla którym wyznaczymy współczynnik korelacji liniowej Pearsona.

```
library(corrplot)
library(olsrr)

korelacje <- pracownicy2 %>%
  select(-c(plec, kat_pracownika, zwiazki)) %>%
  cor()

corrplot(korelacje, method = "number", type = "upper")
```



Możemy zauważyć, że wartości bieżącego wynagrodzenia są najsilniej skorelowane w wartościami wynagrodzenia początkowego. Także doświadczenie i wiek są silnie ze sobą związane, co może sugerować, że obie zmienne wnoszą do modelu podobną informację.

W związku z tym powinniśmy wyeliminować niektóre zmienne z modelu pozostawiając te najważniejsze. Wyróżnia się trzy podjeścia do tego zagadnienia:

- ekspercki dobór cech,
- budowa wszystkich możliwych modeli i wybór najlepszego według określonego kryterium,
- regresja krokowa.

W przypadku budowy wszystkich możliwych modeli należy pamiętać o rosnącej wykładniczo liczbie modeli - $2^p - 1$, gdzie p oznacza liczbę zmiennych objaśniających. w rozważanym przypadku liczba modeli wynosi 255.

```
wszystkie_modele <- ols_step_all_possible(model)
```

W uzyskanym zbiorze danych są informacje o numerze modelu, liczbie użytych zmiennych, nazwie tych zmiennych oraz wiele miar jakości. Te, które warto wziąć pod uwagę to przede wszystkim:

- `rsquare` - współczynnik R-kwadrat,
- `aic` - kryterium informacyjne Akaike,
- `msep` - błąd średniokwadratowy predykcji.

Najwyższa wartość współczynnika R^2 związana jest z modelem zawierającym wszystkie dostępne zmienne objaśniające. Jest to pewna niedoskonałość tej miary, która rośnie wraz z liczbą zmiennych w modelu, nawet jeśli te zmienne nie są istotne.

W przypadku kryteriów informacyjnych oraz błędu średniokwadratowego interesują nas jak najmniejsze wartości. Wówczas jako najlepszy należy wskazać model nr 219 zawierający 6 zmiennych objaśniających.

Metodą, która także pozwoli uzyskać optymalny model, ale przy mniejszym obciążeniu obliczeniowym jest regresja krokowa polegająca na krokowym budowaniu modelu.

```
ols_step_both_aic(model)
```

```
## Stepwise Selection Method
## ------------------------
##
## Candidate Terms:
##
## 1 . plec
## 2 . edukacja
## 3 . kat_pracownika
## 4 . pwynagrodzenie
## 5 . staz
## 6 . doswiadczenie
## 7 . zwiazki
## 8 . wiek
##
##
## Variables Entered/Removed:
##
## - pwynagrodzenie added
## - kat_pracownika added
## - doswiadczenie added
## - staz added
## - edukacja added
## - plec added
##
## No more variables to be added or removed.

##
##
##                                          Stepwise Summary
## ---------------------------------------------------------------------------------------------------
## Variable            Method       AIC             RSS               Sum Sq          R-Sq      Adj. R-Sq
## ---------------------------------------------------------------------------------------------------
## pwynagrodzenie    addition    9862.260    31053506813.535    106862706669.340    0.77484      0.7743
## kat_pracownika    addition    9786.152    26215474648.689    111700738834.186    0.80992      0.8087
## doswiadczenie     addition    9743.487    23853248017.651    114062965465.224    0.82705      0.8255
## staz              addition    9719.469    22576592070.620    115339621412.255    0.83630      0.8345
## edukacja          addition    9707.338    21912088629.084    116004124853.791    0.84112      0.8390
## plec              addition    9703.188    21629051655.016    116287161827.859    0.84317      0.8408
## ---------------------------------------------------------------------------------------------------
```

Otrzymany w ten sposób model jest tożsamy z modelem charakteryzującym się najlepszymi miarami jakości spośród zbioru wszystkich możliwych modeli:

```
wybrany_model <- lm(bwynagrodzenie ~ pwynagrodzenie + kat_pracownika + doswiadczenie + staz + plec + edu
summary(wybrany_model)
```

```
##
## Call:
## lm(formula = bwynagrodzenie ~ pwynagrodzenie + kat_pracownika +
##     doswiadczenie + staz + plec + edukacja, data = pracownicy2)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -22922  -3300   -673   2537  46524
##
## Coefficients:
##                  Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     -6547.147   3402.860  -1.924              0.05496 .
## pwynagrodzenie      1.342      0.073  18.382 < 0.0000000000000002 ***
## kat_pracownika2  6734.992   1631.122   4.129   0.0000431843301918 ***
## kat_pracownika3 11226.635   1368.413   8.204   0.0000000000000023 ***
## doswiadczenie     -22.302      3.571  -6.246   0.0000000009514655 ***
## staz              147.865     31.461   4.700   0.0000034337703087 ***
## plec1           -1878.949    761.703  -2.467              0.01399 *
## edukacja          501.391    160.270   3.128              0.00187 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6820 on 465 degrees of freedom
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8408
## F-statistic: 357.1 on 7 and 465 DF,  p-value: < 0.00000000000000022
```

Uzyskany model charakteryzuje się mniejszym błędem standardowym od modelu ze wszystkimi zmiennymi i tylko jedną nieistotną zmienną. Wyraz wolny (Intercept) nie musi być istotny w modelu.

Wróćmy jeszcze na chwilę do tematu współliniowości zmiennych objaśniających:

```
ols_vif_tol(wybrany_model)
```

```
## # A tibble: 7 x 3
##   Variables       Tolerance   VIF
##   <chr>               <dbl> <dbl>
## 1 pwynagrodzenie      0.298  3.36
## 2 kat_pracownika2     0.687  1.46
## 3 kat_pracownika3     0.360  2.78
## 4 doswiadczenie       0.705  1.42
## 5 staz                0.986  1.01
## 6 plec1               0.683  1.46
## 7 edukacja            0.461  2.17
```

Współczynnik tolerancji wskazuje na procent niewyjaśnionej zmienności danej zmiennej przez pozostałe zmienne objaśniające. Przykładowo współcznnik tolerancji dla początkowego wynagrodzenia wynosi 0,3371, co oznacza, że 33% zmienności początkowego wynagrodzenia nie jest wyjaśnione przez pozostałe zmienne w modelu. Z kolei współczynnik VIF jest obliczany na podstawie wartości współczynnika tolerancji i wskazuje o ile wariancja szacowanego współcznnika regresji jest podwyższona z powodu współliniowości danej zmiennej objaśniającej z pozostałymi zmiennymi objaśniającymi. Wartość współczynnika VIF powyżej 4 należy uznać za wskazującą na współliniowość. W analizowanym przypadku takie zmienne nie występują.

Ocena siły wpływu poszczególnych zmiennych objaśniających na zmienną objaśnianą w oryginalnej postaci modelu nie jest możliwa. Należy wyznaczyć standaryzowane współczynniki beta, które wyliczane są na danych standaryzowanych, czyli takich, które są pozbawione jednostek i cechują się średnią równą 0, a odchyleniem standardowym równym 1. Standaryzacja ma sens tylko dla cech numerycznych, w związku z czym korzystamy z funkcji `mutate_if()`, która jako pierwszy argument przyjmuje warunek, który ma być spełniony, aby była zastosowane przekształcenie podawane jako drugi argument.

```
pracownicy2_std <- pracownicy2 %>%
  mutate_if(is.numeric, funs(scale))
```

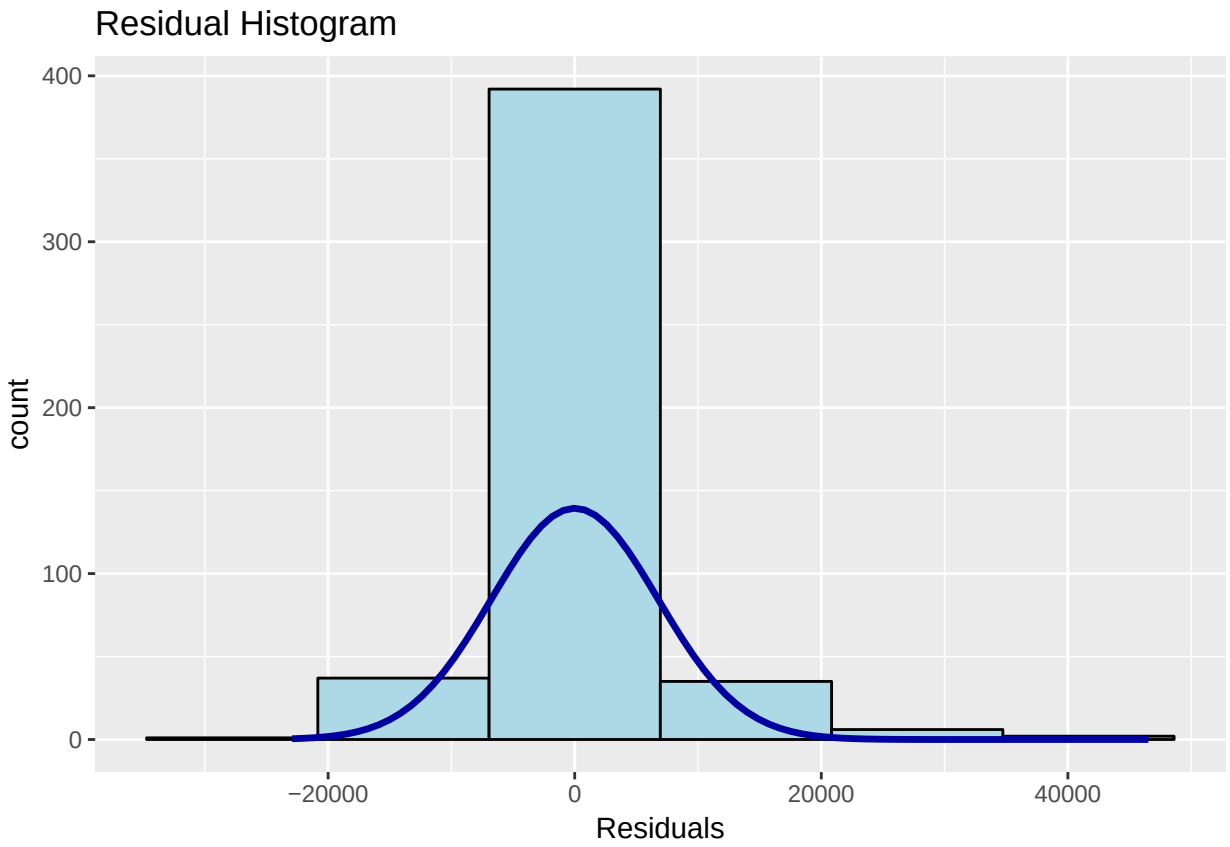
```
wybrany_model_std <- lm(bwynagrodzenie ~ pwynagrodzenie + kat_pracownika +
                        doswiadczenie + staz + plec + edukacja, data = pracownicy2_std)
summary(wybrany_model_std)
```

```
##
## Call:
## lm(formula = bwynagrodzenie ~ pwynagrodzenie + kat_pracownika +
##     doswiadczenie + staz + plec + edukacja, data = pracownicy2_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34098 -0.19306 -0.03939  0.14841  2.72171
##
## Coefficients:
##                 Estimate Std. Error t value             Pr(>|t|)
## (Intercept)     -0.08893    0.03144  -2.828              0.00488 **
## pwynagrodzenie   0.61842    0.03364  18.382 < 0.0000000000000002 ***
## kat_pracownika2  0.39400    0.09542   4.129   0.0000431843301918 ***
## kat_pracownika3  0.65677    0.08005   8.204   0.0000000000000023 ***
## doswiadczenie   -0.13657    0.02187  -6.246   0.0000000009514655 ***
## staz             0.08691    0.01849   4.700   0.0000034337703087 ***
## plec1           -0.10992    0.04456  -2.467              0.01399 *
## edukacja         0.08464    0.02706   3.128              0.00187 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.399 on 465 degrees of freedom
## Multiple R-squared:  0.8432, Adjusted R-squared:  0.8408
## F-statistic: 357.1 on 7 and 465 DF,  p-value: < 0.00000000000000022
```

Spośród cech ilościowych największy wpływ na zmienną objaśnianą mają wartości wynagrodzenia początkowego, staż, edukacja i na końcu doświadczenie.

Reszty czyli różnice pomiędzy obserwowanymi wartościami zmiennej objaśnianej, a wartościami wynikającymi z modelu w klasycznej metodzie najmniejszych kwadratów powinny być zbliżone do rozkładu normalnego. Oznacza to, że najwięcej reszt powinno skupiać się wokół zerowych różnic, natomiast jak najmniej powinno być wartości modelowych znacznie różniących się od tych rzeczywistych.

```
ols_plot_resid_hist(wybrany_model)
```



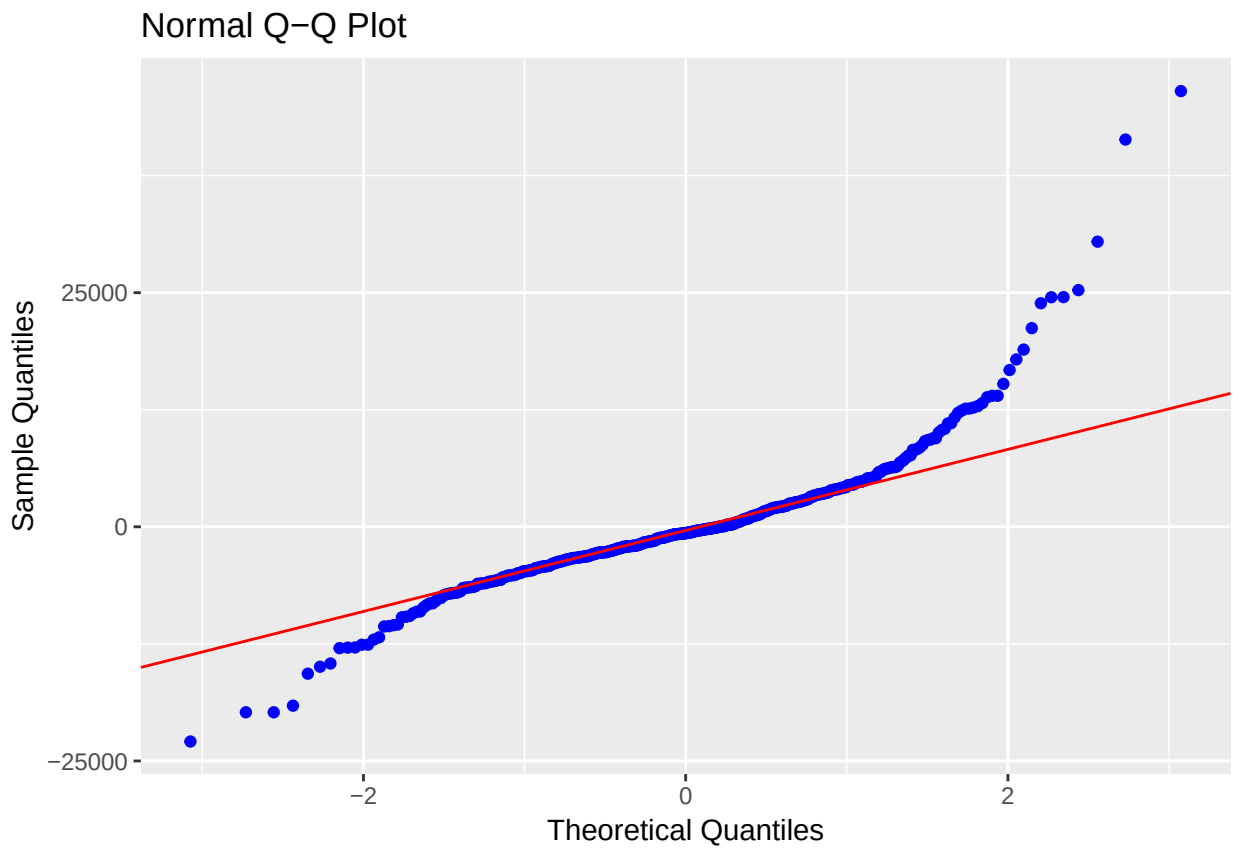
Reszty w naszym modelu wydają się być zbliżone do rozkładu normalnego. Jednoznaczą odpowiedź da jednak odpowiedni test.

```
ols_test_normality(wybrany_model)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.868          0.0000
## Kolmogorov-Smirnov        0.1092         0.0000
## Cramer-von Mises         42.5504         0.0001
## Anderson-Darling         13.0233         0.0000
## -----------------------------------------------
```

Hipoteza zerowa w tych testach mówi o zgodności rozkładu reszt z rozkładem normalnym. Na podstawie wartości p, które są mniejsze od $\alpha = 0,05$ stwierdzamy, że są podstawy do odrzucenia tej hipotezy czyli reszty z naszego modelu nie mają rozkładu normalnego. W diagnostyce przyczyn takiego stanu rzeczy pomoże nam wykres kwantyl-kwantyl:

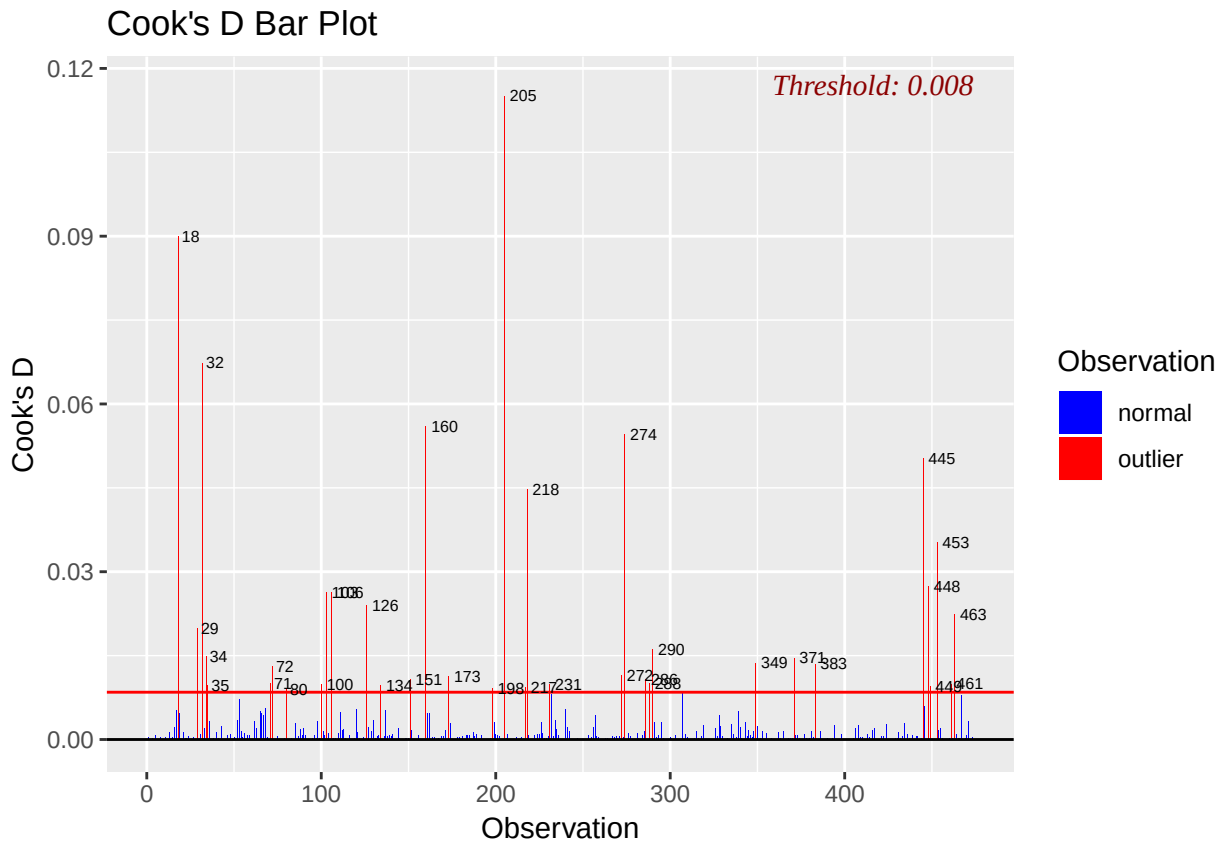
```
ols_plot_resid_qq(wybrany_model)
```



Gdyby wszystkie punkty leżały na prostej to oznaczałoby to normalność rozkładu reszt. Tymczasem po lewej i prawej stronie tego wykresu znajdują się potencjalne wartości odstające, które znacznie wpływają na rozkład reszt modelu.

Wartości odstające można ustalić na podstawie wielu kryteriów. Do jednych z najbardziej popularnych należy odległość Cooka:

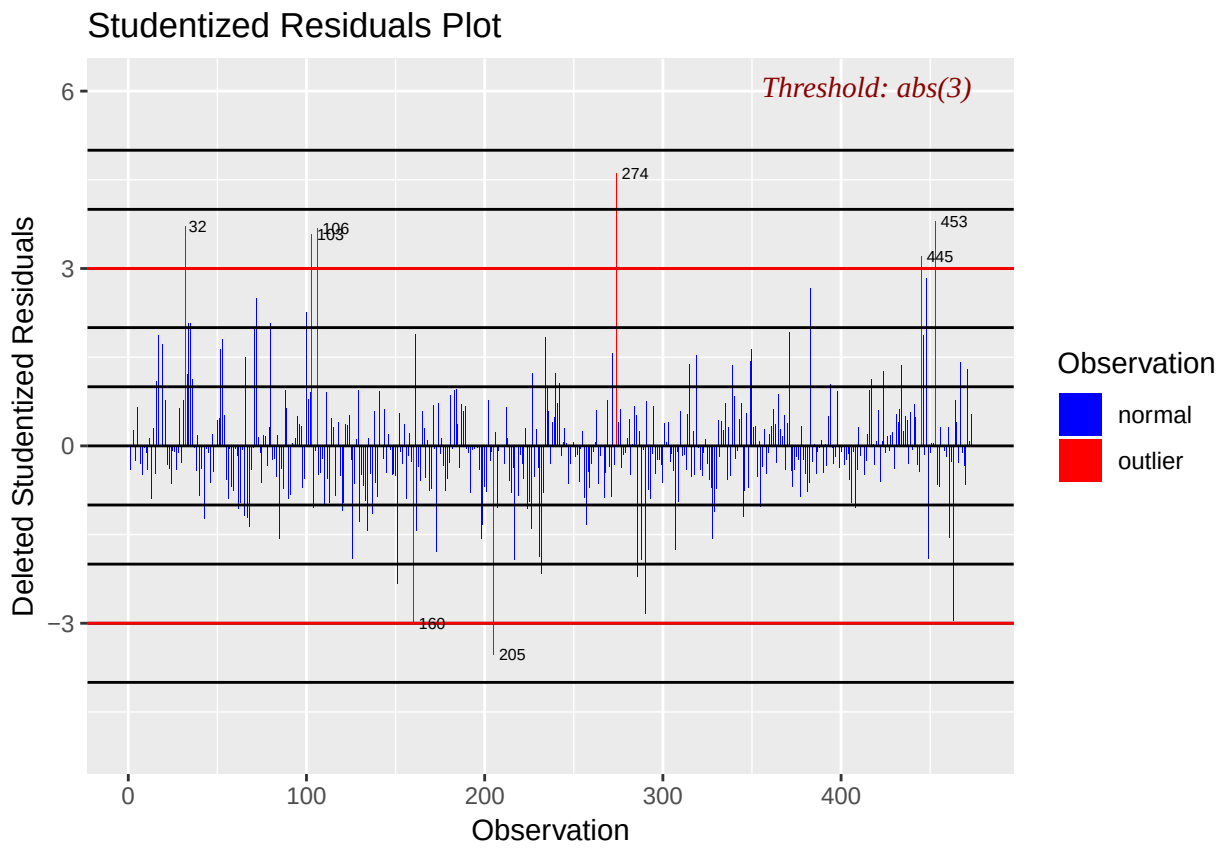
```
cook <- ols_plot_cooksd_bar(wybrany_model)
```



Przypisanie tej funkcji do obiektu zwraca nam tabelę z numerami zidentyfikowanych obserwacji wpływowych. W przypadku odległości Cooka jest to 12 obserwacji.

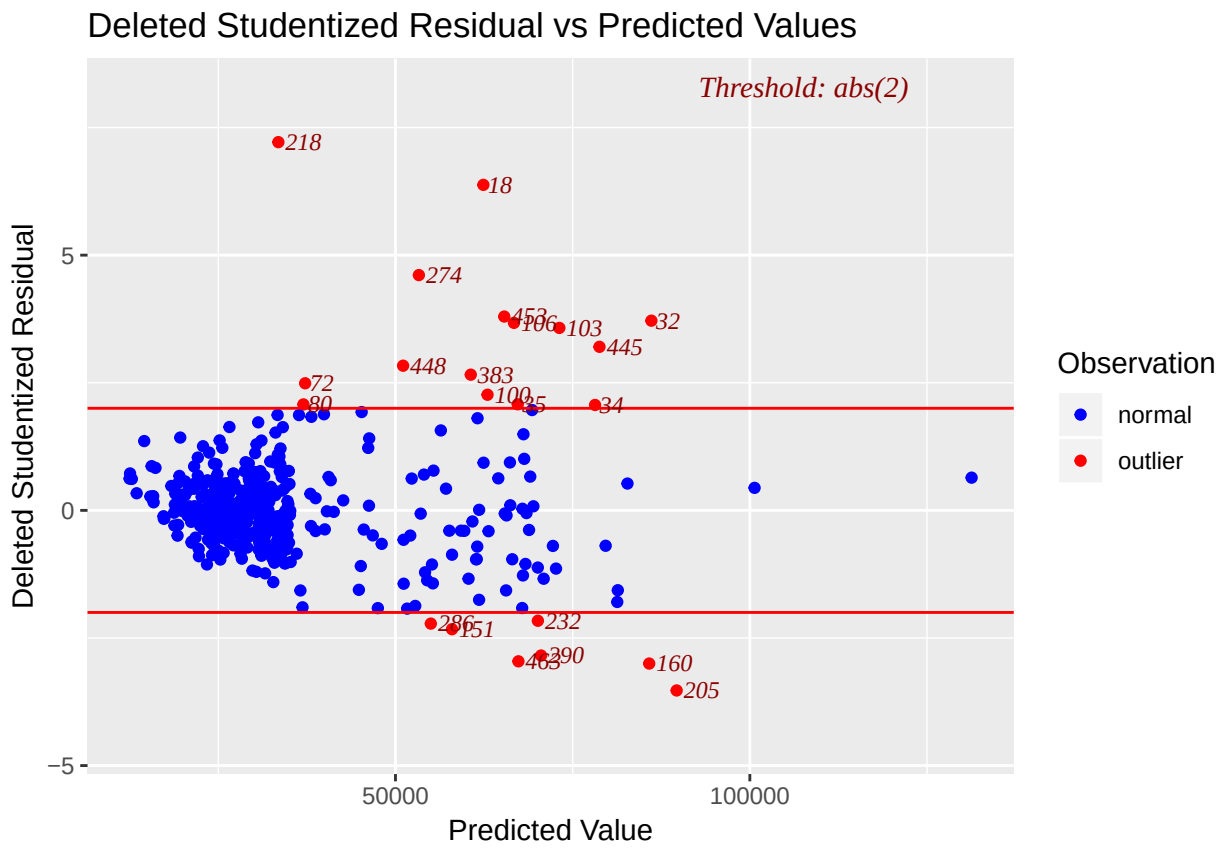
Inną miarą są reszty studentyzowane.

```
stud3 <- ols_plot_resid_stud(wybrany_model)
```



Wyżej wykorzystana funkcja jako kryterium odstawania przyjmuje wartość 3 identyfikując 4 obserwacje wpływowe. Z kolei dodanie do powyższej funkcji przyrostka *fit* powoduje przyjęcie jako granicy wartości równej 2.

```
obs_wplyw <- ols_plot_resid_stud_fit(wybrany_model)
```



W ten sposób zostało zidentyfikowanych 10 obserwacji odstających. Korzystając z tego ostatniego podejścia wyeliminujemy obserwacje odstające ze zbioru uczącego:

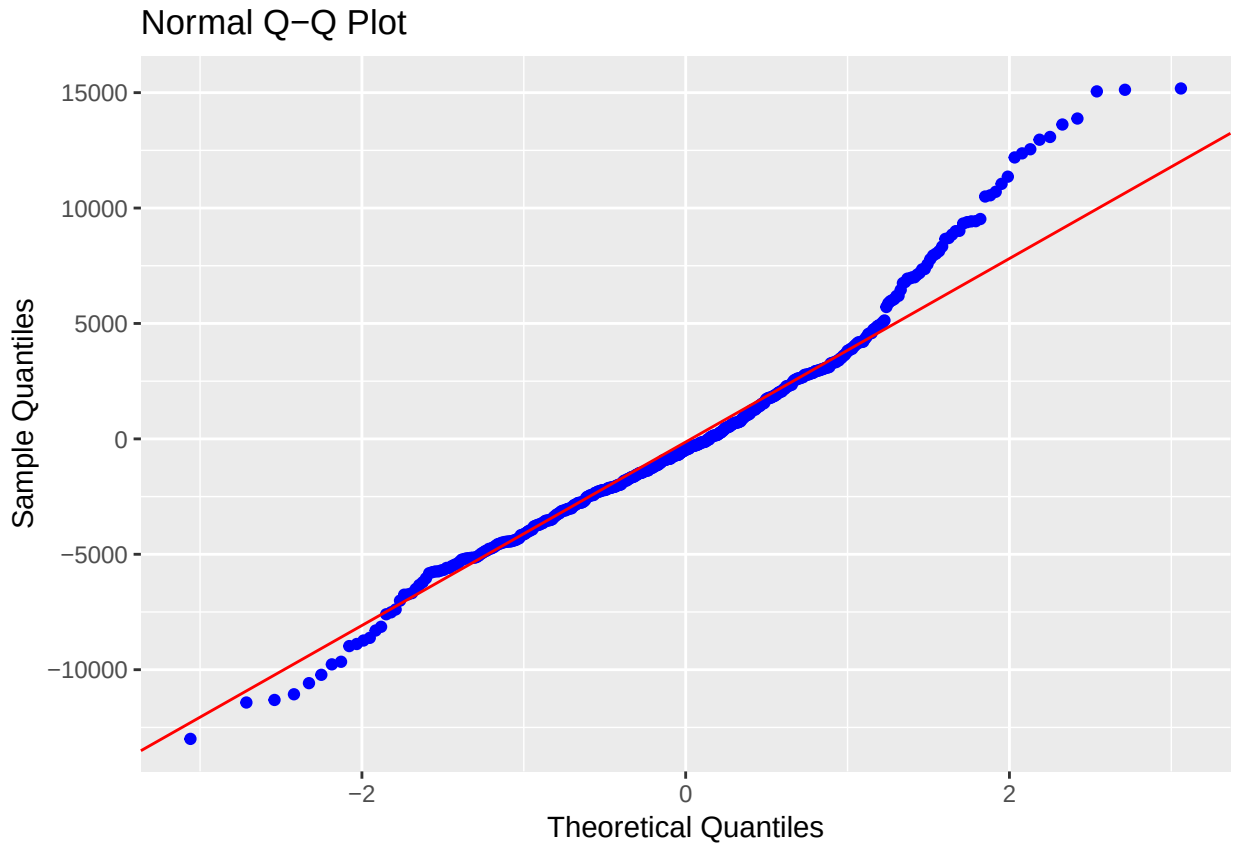
```
nr_obs_wplyw <- obs_wplyw$outliers$observation

bez_obs_wplyw <- pracownicy2[-nr_obs_wplyw,]

wybrany_model_out <- lm(bwynagrodzenie ~ pwynagrodzenie + kat_pracownika + doswiadczenie + staz + plec +
                        data = bez_obs_wplyw)
summary(wybrany_model_out)
```

```
##
## Call:
## lm(formula = bwynagrodzenie ~ pwynagrodzenie + kat_pracownika +
##     doswiadczenie + staz + plec + edukacja, data = bez_obs_wplyw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12997.6  -2816.4   -481.4   2544.6  15180.2
##
## Coefficients:
##                    Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)     -4307.25866  2381.76657  -1.808             0.071217 .
## pwynagrodzenie      1.39451     0.05382  25.908 < 0.0000000000000002 ***
## kat_pracownika2  6097.12115  1102.15833   5.532   0.0000000542431348 ***
## kat_pracownika3  9129.16469   972.97899   9.383 < 0.0000000000000002 ***
## doswiadczenie     -18.87447     2.41685  -7.810   0.0000000000000419 ***
```

```
## staz              120.56289    21.88184   5.510   0.0000000610683928 ***
## plec1           -1483.93151   521.34920  -2.846              0.004628 **
## edukacja          384.58159   109.81914   3.502              0.000509 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4592 on 443 degrees of freedom
## Multiple R-squared:  0.8986, Adjusted R-squared:  0.897
## F-statistic:   561 on 7 and 443 DF,  p-value: < 0.00000000000000022
```

Model dopasowny na takim zbiorze charakteryzuje się dużo mniejszym błędem standardowym oraz wyższym współczynnikiem R^2. Sprawdźmy w takim razie normalność reszt.

```
ols_plot_resid_qq(wybrany_model_out)
```



Wykres kwantyl-kwantyl wygląda już dużo lepiej, ale dla pewności przeprowadzimy testy statystyczne.

```
ols_test_normality(wybrany_model_out)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9708         0.0000
## Kolmogorov-Smirnov        0.0675         0.0328
## Cramer-von Mises         39.1404         0.0001
## Anderson-Darling          3.8103         0.0000
## -----------------------------------------------
```

Tylko jeden test wskazał zgodność rozkładu reszt z rozkładem normalnym.

3.2.1 Zadanie

Na podstawie zbioru dotyczącego 50 startupów określ jakie czynniki w największym stopniu wpływają na przychód startupów.

```
startupy <- read.csv("data/50_Startups.csv")

summary(startupy)
```

```
##    R.D.Spend      Administration   Marketing.Spend         State
##  Min.   :     0   Min.   : 51283   Min.   :     0   California:17
##  1st Qu.: 39936   1st Qu.:103731   1st Qu.:129300   Florida   :16
##  Median : 73051   Median :122700   Median :212716   New York  :17
##  Mean   : 73722   Mean   :121345   Mean   :211025
##  3rd Qu.:101603   3rd Qu.:144842   3rd Qu.:299469
##  Max.   :165349   Max.   :182646   Max.   :471784
##      Profit
##  Min.   : 14681
##  1st Qu.: 90139
##  Median :107978
##  Mean   :112013
##  3rd Qu.:139766
##  Max.   :192262
```

Chapter 4

Grupowanie

Metody grupowania są wykorzystywane np. do segmentacji klientów, w przypadku, gdy nie jest znany końcowy podział.

4.1 Metoda k-średnich

Algorytm:

1. Wskaź liczbę grup k.
2. Wybierz dowolne k punktów jako centra grup.
3. Przypisz każdą z obserwacji do najbliższego centroidu.
4. Oblicz nowe centrum grupy.
5. Przypisz każdą z obserwacji do nowych centroidów. Jeśli któraś obserwacja zmieniła grupę - przejdź do kroku nr 4, a w przeciwnym przypadku zakończ algorytm.

Zalety:

- dobrze działa zarówno na małych, jak i dużych zbiorach
- efektywny

Wady:

- trzeba wskazać liczbę grup
- losowy wybór punktów początkowych

4.2 Metoda hierarchiczna

Algorytm:

1. Każda obserwacji stanowi jedną z N pojedyńczych grup.
2. Na podstawie macierzy odległości połącz dwie najbliżej leżące obserwacje w jedną grupę ($N-1$ grup).
3. Połącz dwa najbliżej siebie leżące grupy w jedną ($N-2$ grup).
4. Powtórz krok nr 3, aż do uzyskania jednej grupy.

Zalety:

- prosty sposób ustalenia liczby grup
- praktyczny sposób wizualizacji

Wady:

- nieodpowiedni dla dużych zbiorów

4.2.1 Zadanie

Na podstawie zbioru zawierającego informacje o klientach sklepu dokonaj grupowania klientów.

Opis zbioru:

- klientID - identyfikator klienta
- plec - płeć
- wiek - wiek
- roczny_dochod - roczny dochód wyrażony w tys. dolarów
- wskaznik_wydatkow - klasyfikacja sklepu od 1 do 100

grupowanie.Rmd

cały projekt

4.2.2 Zadanie 2

Dokonaj grupowania danych dotyczących 32 samochodów według następujących zmiennych: pojemność, przebieg, lata oraz cena.

4.2.3 Zadanie 3

Rozpoznawanie czynności na podstawie danych z przyspieszeniomierza w telefonie: User Identification From Walking Activity Data Set

Chapter 5

Klasyfikacja

A visual introduction to machine learning - niestety powstała tylko jedna część.

5.1 Drzewa klasyfikacyjne

Zalety:

- łatwa interpretacja
- nie trzeba normalizować cech
- rozwiązuje problemy liniowe i nieliniowe

Wady:

- mała efektywność przy małych zbiorach danych
- łatwo można przeuczyć

5.2 KNN

Algorytm:

1. Określ liczbę sąsiadów - K
2. Wyznacz K sąsiadów dla nowego punktu na podstawie wybranej odległości
3. Oblicz liczbę sąsiadów, w każdej z grup
4. Przypisz nową obserwację do grupy, w której ma więcej najbliższych sąsiadów

Zalety:

- łatwa interpretacja
- szybki i efektywny

Wady:

- trzeba określić liczbę sąsiadów

5.2.1 Zadanie

Zbuduj model klasyfikacyjny dla zbioru danych dotyczących cech internautów oraz informacji czy zamówili reklamowany produkt czy nie.

Przeprowadź imputację braków danych dla zbioru pracowników.

Chapter 6

Materiały z zajęć

6.1 28.10.2018

Wprowadzenie do R

Analiza sejmików

6.2 18.11.2018

Analiza struktury

Rossmann

Analiza struktury w R

6.3 16.12.2018

Prezentacja

Pracownicy

Korelacja w R

Regresja w R

6.4 26.01.2019

Pensja i doświaczenie

Pracownicy

Opis zbioru:

- id - kod pracownika
- plec - płeć pracownika (0 - mężczyzna, 1 - kobieta)
- data_urodz - data urodzenia
- edukacja - wykształcenie (w latach nauki)
- kat_pracownika - grupa pracownicza (1 - ochroniarz, 2 - urzędnik, 3 - menedżer)

- bwynagrodzenie - bieżące wynagrodzenie
- pwynagrodzenie - początkowe wynagrodzenie
- staz - staż pracy (w miesiącach)
- doswiadczenie - poprzednie zatrudnienie (w miesiącach)
- zwiazki - przynależność do związków zawodowych (0 - nie, 1 - tak)
- wiek - wiek (w latach)

Regresja w R