Zastosowanie Regresji Liniowej

Wojciech Kosztyła, Wojciech Węgrzynek

Zastosowanie regresji liniowej

Należy dokonać analizy wybranego zbioru danych metodą regresji liniowej (wielokrotnej).W szczególności należy:

- ocenić czy którykolwiek z predyktorów ma istotny wpływ na odpowiedź;
- (jeśli odpowiedź w poprzednim punkcie jest pozytywna) dla każdego predyktora osobno ocenić istotność jego wpływu na odpowiedź;
- przeanalizować charakter wpływu każdego predyktora osobno na odpowiedź (dodatni/ujemny, duży/mały);
- ocenić dopasowanie modelu do danych.

Projekt wykonujemy na platformie Google Collab, co umożliwia nam łatwą pracę na tych samych danych wejściowych i wyjściowych.
Nasz dataset to “US Census Demographic Data” (https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data). Przed regresją liniową zrobiliśmy preprocessing, a w jego skład wchodziło m.in. wyrzucenie kolumn z błędami, kolumn bez wartości liczbowych i przeskalowanie.

```
[130]  1 fit_all <- lm(Income ~ ., data = data_for_regression)
       2 summary(fit_all)
```

```
Call:
lm(formula = Income ~ ., data = data_for_regression)

Residuals:
   Min     1Q Median     3Q    Max
-17545  -2717   -336   2350  33688

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.749e+04  2.417e+05   0.279   0.7801
TotalPop     -2.191e-03  3.364e-04  -6.512 8.57e-11 ***
Men           4.957e+04  4.417e+03  11.224  < 2e-16 ***
Women                NA         NA      NA       NA
Hispanic     -1.009e+02  6.101e+01  -1.653   0.0983 .
White        -1.577e+02  6.167e+01  -2.557   0.0106 *
Black        -1.538e+02  6.136e+01  -2.507   0.0122 *
Native       -2.508e+01  6.706e+01  -0.374   0.7084
Asian         5.460e+02  8.651e+01   6.312 3.14e-10 ***
Pacific      -9.256e+02  3.006e+02  -3.079   0.0021 **
Citizen      -3.357e+04  2.516e+03 -13.344  < 2e-16 ***
Poverty      -9.204e+02  3.553e+01 -25.907  < 2e-16 ***
ChildPoverty  3.243e+01  2.341e+01   1.385   0.1660
Professional  1.690e+03  1.370e+03   1.234   0.2173
Service       1.084e+03  1.370e+03   0.792   0.4286
Office        1.211e+03  1.370e+03   0.883   0.3770
Construction  1.205e+03  1.370e+03   0.880   0.3790
Production    1.062e+03  1.370e+03   0.775   0.4383
Drive        -3.527e+02  1.263e+03  -0.279   0.7801
Carpool      -5.000e+02  1.264e+03  -0.396   0.6924
Transit      -4.036e+02  1.263e+03  -0.320   0.7493
Walk         -3.922e+02  1.263e+03  -0.310   0.7562
OtherTransp  -3.306e+02  1.263e+03  -0.262   0.7935
WorkAtHome   -3.220e+02  1.264e+03  -0.255   0.7989
MeanCommute   4.297e+02  2.010e+01  21.374  < 2e-16 ***
Employed      4.861e+04  2.647e+03  18.364  < 2e-16 ***
PrivateWork  -1.141e+03  1.540e+03  -0.741   0.4588
PublicWork   -1.175e+03  1.540e+03  -0.763   0.4456
SelfEmployed -1.603e+03  1.539e+03  -1.041   0.2979
FamilyWork   -1.531e+03  1.545e+03  -0.991   0.3219
Unemployment  1.746e+02  3.744e+01   4.664 3.23e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4932 on 3188 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8553,    Adjusted R-squared:  0.854
F-statistic:   650 on 29 and 3188 DF,  p-value: < 2.2e-16
```

Predykatory o istotnych wpływach:

- TotalPop - bardzo duży wpływ
- Men - bardzo duży wpływ
- Hispanic - bardzo mały wpływ
- White - mały wpływ
- Black - mały wpływ
- Asian - bardzo duży wpływ
- Pacific - duży wpływ
- Citizen - bardzo duży wpływ
- Poverty - bardzo duży wpływ
- MeanCommute - bardzo duży wpływ
- Employed - bardzo duży wpływ
- Unemployment - mniejszy wpływ

Wpływy predykatorów:

- TotalPop - ujemny duży
- Men - dodatni bardzo duży
- Hispanic - ujemny mały
- White - ujemny mały
- Black - ujemny mały
- Native - ujemny bardzo mały
- Asian - dodatni mały
- Pacific - ujemny mały
- Citizen - ujemny bardzo duży
- Poverty - ujemny mały
- ChildPoverty - dodatni bardzo mały
- Professional - dodatni duży
- Service - dodatni duży
- Office - dodatni duży
- Construction - dodatni duży
- Production - dodatni duży
- Drive - ujemny mały
- Carpool - ujemny mały
- Transit - ujemny mały
- Walk - ujemny mały
- OtherTransp - ujemny mały
- WorkAtHome - ujemny mały
- MeanCommute - dodatni mały
- Employed - dodatni bardzo duży
- PrivateWork - ujemny duży
- PublicWork - ujemny duży
- SelfEmployed - ujemny duży
- FamilyWork - ujemny duży
- Unemployment - dodatni mały

Dopasowanie modelu do danych muszę ocenić negatywnie. Wynik R-squared wynisi jedynie 0.8553. F-statistic jest też niezadowalające.