multiple linear regression

Katarzyna Tokarczuk

2024-05-08

Task 1

```
dane = read.delim("zanieczyszczenia.txt")
head(dane)
```

```
##   month day hour PM2.5 PM10 SO2 NO2   CO O3 TEMP
## 1     3   1   12     8    8   3  12  300 64  2.8
## 2     3   2   12    19   22  17  45  600 37  2.1
## 3     3   3   12    95  106  60  73 1700 28 11.8
## 4     3   4   12     3    7   5  15  500 83 13.4
## 5     3   5   12   127  147  76  68 1800 33 12.3
## 6     3   6   12   160  174 103  87 2700 61  9.7
```

Task 2

```
dane <- na.omit(dane)
```

Task 3

```
correlation_matrix <- cor(dane)
```

```
## Warning in cor(dane): odchylenie standardowe wynosi zero
```

```
print(correlation_matrix)
```

```
##              month         day hour       PM2.5         PM10          SO2
## month  1.000000000 0.023617409   NA -0.03905050 -0.028940508 -0.203277620
## day    0.023617409 1.000000000   NA  0.02163491  0.039662520  0.005877225
## hour            NA          NA    1          NA           NA           NA
## PM2.5 -0.039050498 0.021634905   NA  1.00000000  0.929209653  0.542740337
## PM10  -0.028940508 0.039662520   NA  0.92920965  1.000000000  0.559675287
## SO2   -0.203277620 0.005877225   NA  0.54274034  0.559675287  1.000000000
## NO2    0.001417258 0.025656777   NA  0.74143450  0.745182665  0.657386651
## CO     0.013335455 0.005050268   NA  0.80805436  0.769386013  0.603814813
## O3    -0.125256657 0.014566740   NA -0.09498647 -0.101289135 -0.295587541
```

```
## TEMP   0.143772474 0.012860321   NA -0.02870710 -0.007391401 -0.286556128
##                NO2           CO          O3         TEMP
## month  0.001417258  0.013335455 -0.12525666  0.143772474
## day    0.025656777  0.005050268  0.01456674  0.012860321
## hour            NA           NA          NA           NA
## PM2.5  0.741434501  0.808054364 -0.09498647 -0.028707098
## PM10   0.745182665  0.769386013 -0.10128914 -0.007391401
## SO2    0.657386651  0.603814813 -0.29558754 -0.286556128
## NO2    1.000000000  0.759606983 -0.43586431 -0.237908000
## CO     0.759606983  1.000000000 -0.29613393 -0.194807142
## O3    -0.435864309 -0.296133926  1.00000000  0.681249890
## TEMP  -0.237908000 -0.194807142  0.68124989  1.000000000
```

```
dane <- dane[, !(names(dane) %in% c("hour"))]
```

Task 4

```
model <- lm(TEMP ~ SO2 + NO2 + CO + O3, data = dane)
#excluding Month and Day, and omitting PM2.5 and PM10 due to high correlation.
```

Task 5

```
#a)
wspolczynniki <- coef(model)
print(wspolczynniki)
```

```
##   (Intercept)           SO2           NO2            CO            O3
##  2.5948689246 -0.1041748470  0.1027068109 -0.0004575303  0.1641486729
```

```
#b)
RSS <- sum(model$residuals^2)
RSE <- sqrt(RSS / (length(model$residuals) - length(wspolczynniki) - 1))
R_squared <- summary(model)$r.squared

RSS
```

```
## [1] 79926.92
```

```
RSE
```

```
## [1] 7.933132
```

```
R_squared
```

```
## [1] 0.4960048
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = TEMP ~ SO2 + NO2 + CO + O3, data = dane)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.2992  -5.8787   0.4067   5.8046  20.5383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.5948689  0.6359766   4.080 4.78e-05 ***
## SO2         -0.1041748  0.0133929  -7.778 1.51e-14 ***
## NO2          0.1027068  0.0148171   6.932 6.60e-12 ***
## CO          -0.0004575  0.0003574  -1.280    0.201
## O3           0.1641487  0.0051091  32.129  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.93 on 1271 degrees of freedom
## Multiple R-squared:  0.496,  Adjusted R-squared:  0.4944
## F-statistic: 312.7 on 4 and 1271 DF,  p-value: < 2.2e-16
```

Task 6

```r
nowe_dane <- data.frame(SO2 = 25, NO2 = 90, CO = 2000, O3 = 50)
predykcja <- predict(model, newdata = nowe_dane)
print(predykcja)
```

```
##        1
## 16.52648
```