大數據統計分析與預測 第十八章作業(Simple Linear Regression)

Created by Weber YC, Huang (黃彥鈞 m946108006) 2019-12-23

Data : CVD_ALL.csv

Table in need : 64489 × 2 (sample head(6))

WAIST	抽菸量
81.0	2
79.0	2
86.5	1
84.0	0
96.0	1
94.0	0

Q. 抽菸量預測腰圍 :

> *R - code :*

```
# preprocessing...
# set working dir
setwd('C:/Users/doudi/Downloads')

# load packages
library(dplyr)
library(magrittr)

# load file
data <- read.csv('CVD_ALL.csv', encoding = 'utf-8')
df <- data[,c(6,16)]
colnames(df) <- c('waist', 'Smoke')

# drop NA
sapply(df, function(x) {sum(is.na(x))})
df = df %>% na.omit()

# dummy vars
df$Smoke = df$Smoke %>% as.factor()
justdummy_data <- model.matrix(~df$Smoke-1)
alldummy_data <- cbind(df,justdummy_data)
colnames(alldummy_data) <- c('waist', 'Smoke', 'zero',
'one', 'two', 'three')

# dummy test
alldummy_data.fit <- lm(waist ~ zero + one + two + three,
data = alldummy_data)
```

```
summary(alldummy_data.fit)

# no dummy
df.fit <- lm(waist ~ Smoke, data = df)
summary(df.fit)
```

1. 類別化(Dummy variables)

Run code ...

```
Call:
lm(formula = waist ~ zero + one + two + three, data =
alldummy_data)

Residuals:
    Min      1Q  Median      3Q     Max
-46.488  -7.792  -0.792   7.208 102.208

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  87.4878     0.8076 108.327  < 2e-16 ***
zero        -10.6960     0.8091 -13.220  < 2e-16 ***
one          -5.3467     0.8123  -6.582 4.67e-11 ***
two          -1.8530     0.8487  -2.183    0.029 *
three            NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 10.34 on 61168 degrees of freedom
Multiple R-squared:  0.0591,    Adjusted R-squared:
 0.05905
F-statistic:  1281 on 3 and 61168 DF,  p-value: < 2.2e-16
```

2. 未類別化(No dummy variables)

Run code ...

```
Call:
lm(formula = waist ~ Smoke, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-46.488  -7.792  -0.792   7.208 102.208

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 76.79177    0.04863 1578.99   <2e-16 ***
Smoke1       5.34930    0.09947   53.78   <2e-16 ***
Smoke2       8.84300    0.26527   33.34   <2e-16 ***
```

```
Smoke3      10.69603    0.80909   13.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 10.34 on 61168 degrees of freedom
Multiple R-squared:  0.0591,    Adjusted R-squared:
 0.05905
F-statistic:  1281 on 3 and 61168 DF,  p-value: < 2.2e-16
```

3. 比較 1, 2 結果：

   1, 2 出來結果一樣，用哪一種方法沒有太大的差別。