Modul 08 - Simple Linear Regression

Roni Yunis

3/19/2021

Pengantar

Regresi liner sederhana menggambarkan hubungan liner antara dua variabel, yaitu variabel independen (x) dan variabel dependen (y). $y = b0 + b1 * x$

Nilai b disebut dengan bobot regresi (koefisien beta), digunakan untuk mengukur hubungan antara variabel prediktor dan hasil

Load Packages

```
#Split dataset
library (caTools)
#Predicting result visualization
library(ggplot2)
#Import dataset marketing pada library datarium
library(datarium)
```

```
## Warning: package 'datarium' was built under R version 4.0.5
```

View dataset

```
head(marketing)
```

```
##   youtube facebook newspaper sales
## 1  276.12    45.36     83.04 26.52
## 2   53.40    47.16     54.12 12.48
## 3   20.64    55.08     83.16 11.16
## 4  181.80    49.56     70.20 22.20
## 5  216.96    12.96     70.08 15.48
## 6   10.44    58.68     90.00  8.64
```

Melihat dimensi dari data, dengan menggunakan fungsi dim()

```
dim(marketing)
```

```
## [1] 200   4
```

Melihat struktur dari data, dengan menggunakan fungsi str()

```
str(marketing)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ youtube  : num  276.1 53.4 20.6 181.8 217 ...
##  $ facebook : num  45.4 47.2 55.1 49.6 13 ...
##  $ newspaper: num  83 54.1 83.2 70.2 70.1 ...
##  $ sales    : num  26.5 12.5 11.2 22.2 15.5 ...
```

Exploratory Data Analysis

```
summary(marketing)
```

```
##     youtube          facebook       newspaper          sales      
##  Min.   :  0.84   Min.   : 0.00   Min.   :  0.36   Min.   : 1.92  
##  1st Qu.: 89.25   1st Qu.:11.97   1st Qu.: 15.30   1st Qu.:12.45  
##  Median :179.70   Median :27.48   Median : 30.90   Median :15.48  
##  Mean   :176.45   Mean   :27.92   Mean   : 36.66   Mean   :16.83  
##  3rd Qu.:262.59   3rd Qu.:43.83   3rd Qu.: 54.12   3rd Qu.:20.88  
##  Max.   :355.68   Max.   :59.52   Max.   :136.80   Max.   :32.40
```

```
cor(marketing)
```

```
##              youtube   facebook  newspaper     sales
## youtube   1.00000000 0.05480866 0.05664787 0.7822244
## facebook  0.05480866 1.00000000 0.35410375 0.5762226
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## sales     0.78222442 0.57622257 0.22829903 1.0000000
```

Sekarang kita akan mencoba melihat korelasi antara facebook dengan sales

```
#menghitung korelasi antar variabel
korfacebook <- cor(marketing$facebook, marketing$sales)
korfacebook
```

```
## [1] 0.5762226
```

Koefisien korelasi mengukur tingkat hubungan antara dua variabel x dan y. Nilainya berkisar antara -1 (korelasi negatif sempurna: ketika x meningkat, y menurun) dan +1 (korelasi positif sempurna: ketika x meningkat, y meningkat).

Nilai yang mendekati 0 menunjukkan hubungan yang lemah antara variabel. Korelasi yang rendah (-0,2 <x <0,2) mungkin menunjukkan bahwa banyak variasi dari variabel hasil (y) tidak dijelaskan oleh prediktor (x). Dalam kasus seperti itu, kita mungkin harus mencari variabel prediktor yang lebih baik.

Dalam contoh ini, koefisien korelasinya antara variabel facebook dan sales adalah sebesar 0,57

Bagi dataset kedalam data training dan data testing

```
splitdata <- sample.split(marketing$sales, SplitRatio = 0.7)
trainingset <- subset(marketing, splitdata == TRUE)
testingset <- subset(marketing, splitdata == FALSE)
```

```
dim(trainingset)
```

```
## [1] 140   4
```

```
dim(testingset)
```

```
## [1] 60  4
```

Bisa dilihat bahwa untuk data training ada 140 baris data, dan untuk data testing ada 60 baris data

Model Regresi Sederhana

liner regresi sederhana pada data training

```
lm.r <- lm(sales ~ facebook,
           data = trainingset)
summary(lm.r)
```

```
##
## Call:
## lm(formula = sales ~ facebook, data = trainingset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2346  -2.7615   0.8173   3.7172  10.1028
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.03103    0.88638  12.445  < 2e-16 ***
## facebook     0.19199    0.02566   7.481 7.81e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.423 on 138 degrees of freedom
## Multiple R-squared:  0.2885, Adjusted R-squared:  0.2834
## F-statistic: 55.96 on 1 and 138 DF,  p-value: 7.813e-12
```

Berdasarkan output diatas bisa dijelaskan bahwa: Persamaan regresi linernya adalah $sales = 11,17 + 0,190 * facebook$

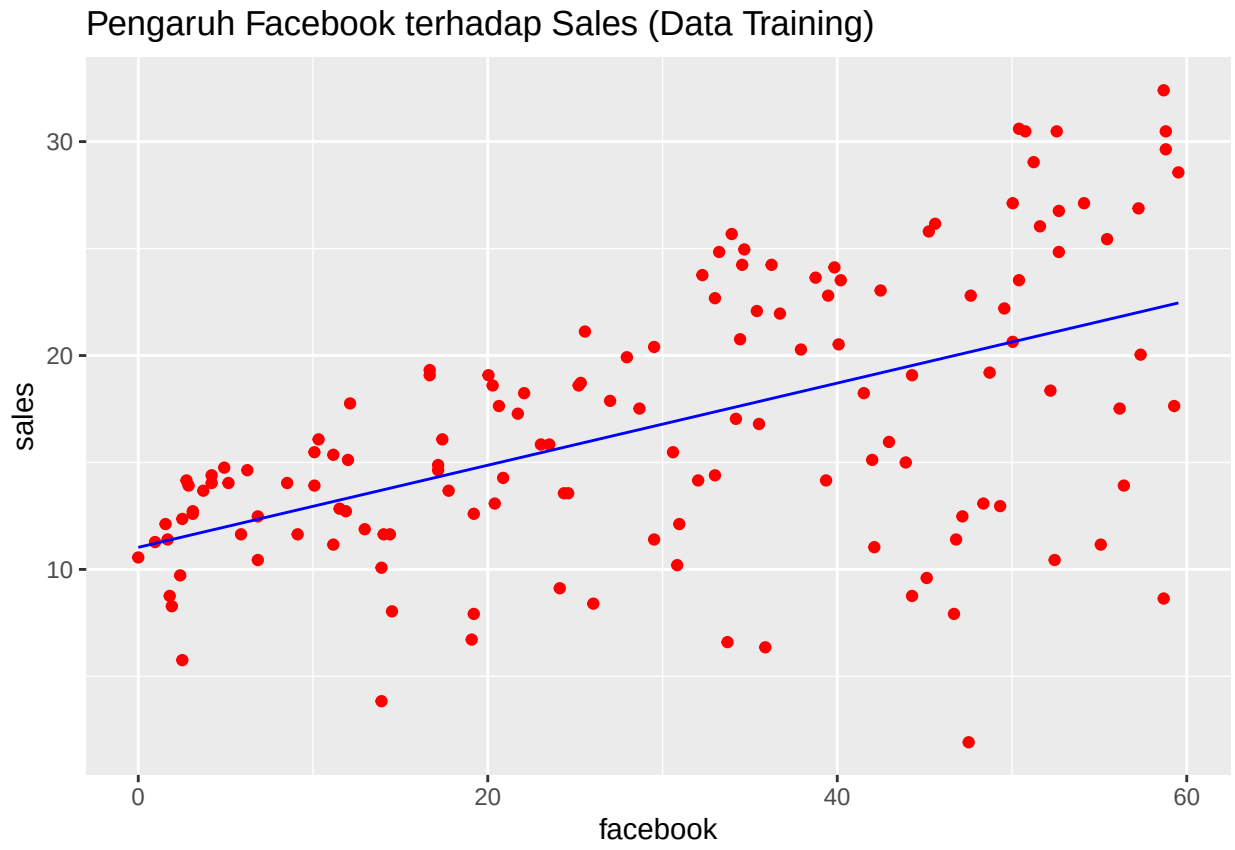
Kalau kita lihat dari model diatas bahwa facebook punya hubungan signifikan terhadap penjualan, artinya nilai penjualan dapat ditingkatkan dari anggaran iklan pada facebook.

Prediksi model regresi dengan data testing

```
yprediksi <- predict(lm.r, newdata = testingset)
```

Visualasi hasil data training

```
ggplot() + geom_point(aes(x = trainingset$facebook,
                          y = trainingset$sales), colour = 'red') +
  geom_line(aes(x = trainingset$facebook,
                y = predict(lm.r, newdata = trainingset)), colour = 'blue') +
  ggtitle('Pengaruh Facebook terhadap Sales (Data Training)') +
  xlab('facebook') +
  ylab('sales')
```



Visualasi hasil data testing

```
ggplot() + geom_point(aes(x = testingset$facebook,
                          y = testingset$sales), colour = 'red') +
  geom_line(aes(x = testingset$facebook,
                y = predict(lm.r, newdata = testingset)), colour = 'blue') +
```

```
ggtitle('Pengaruh Facebook terhadap Sales (Data Testing)') +
xlab('Facebook') +
ylab('Sales')
```