Modul08-Simple Linear Regression

Roni Yunis

3/19/2021

Load Packages

```
#Split dataset
library (caTools)
#Predicting result visualization
library(ggplot2)
#Import format data xlsx
library(readxl)
```

Import dataset

```
credit <- read_excel("data/Credit Risk Data.xlsx", sheet = "Base Data")
head(credit)
```

```
## # A tibble: 6 x 12
##   LoanPurpose Checking Savings MonthsCustomer MonthsEmployed Gender
##   <chr>          <dbl>   <dbl>          <dbl>          <dbl> <chr>
## 1 Small Appl~        0     739             13             12 M
## 2 Furniture          0    1230             25              0 M
## 3 New Car            0     389             19            119 M
## 4 Furniture        638     347             13             14 M
## 5 Education        963    4754             40             45 M
## 6 Furniture       2827       0             11             13 M
## # ... with 6 more variables: MaritalStatus <chr>, Age <dbl>, Housing <chr>,
## #   Years <dbl>, Job <chr>, CreditRisk <chr>
```

Melihat dimensi dari data, dengan menggunakan fungsi dim()

```
dim(credit)
```

```
## [1] 425  12
```

Melihat struktur dari data, dengan menggunakan fungsi str()

```
str(credit)
```

```
## tibble [425 x 12] (S3: tbl_df/tbl/data.frame)
##  $ LoanPurpose   : chr [1:425] "Small Appliance" "Furniture" "New Car" "Furniture" ...
##  $ Checking      : num [1:425] 0 0 0 638 963 ...
##  $ Savings       : num [1:425] 739 1230 389 347 4754 ...
##  $ MonthsCustomer: num [1:425] 13 25 19 13 40 11 13 14 37 25 ...
##  $ MonthsEmployed: num [1:425] 12 0 119 14 45 13 16 2 9 4 ...
##  $ Gender        : chr [1:425] "M" "M" "M" "M" ...
##  $ MaritalStatus : chr [1:425] "Single" "Divorced" "Single" "Single" ...
##  $ Age           : num [1:425] 23 32 38 36 31 25 26 27 25 43 ...
##  $ Housing       : chr [1:425] "Own" "Own" "Own" "Own" ...
##  $ Years         : num [1:425] 3 1 4 2 3 1 3 1 2 1 ...
##  $ Job           : chr [1:425] "Unskilled" "Skilled" "Management" "Unskilled" ...
##  $ CreditRisk    : chr [1:425] "Low" "High" "High" "High" ...
```

#Cek data kosong/missing value

```
summary(credit)
```

```
##  LoanPurpose           Checking        Savings      MonthsCustomer
##  Length:425         Min.   :    0   Min.   :    0   Min.   : 5.0
##  Class :character   1st Qu.:    0   1st Qu.:  228   1st Qu.:13.0
##  Mode  :character   Median :    0   Median :  596   Median :19.0
##                     Mean   : 1048   Mean   : 1813   Mean   :22.9
##                     3rd Qu.:  560   3rd Qu.:  921   3rd Qu.:28.0
##                     Max.   :19812   Max.   :19811   Max.   :73.0
##  MonthsEmployed     Gender          MaritalStatus          Age
##  Min.   :  0.0   Length:425         Length:425         Min.   :18.0
##  1st Qu.:  6.0   Class :character   Class :character   1st Qu.:26.0
##  Median : 20.0   Mode  :character   Mode  :character   Median :32.0
##  Mean   : 31.9                                         Mean   :34.4
##  3rd Qu.: 47.0                                         3rd Qu.:41.0
##  Max.   :119.0                                         Max.   :73.0
##    Housing              Years          Job             CreditRisk
##  Length:425         Min.   :1.00   Length:425         Length:425
##  Class :character   1st Qu.:2.00   Class :character   Class :character
##  Mode  :character   Median :3.00   Mode  :character   Mode  :character
##                     Mean   :2.84
##                     3rd Qu.:4.00
##                     Max.   :4.00
```

Sekarang kita akan mencoba melihat korelasi antara Age (Umur) dengan MonthsEmployed (Bulan Bekerja).

```
#menghitung korelasi antar variabel
korelasi <- cor(credit$Age, credit$MonthsEmployed)
korelasi
```

```
## [1] 0.3067985
```

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 tidak terlalu besar yaitu 0,31, jadi kita bisa melanjutkan dengan membangun model linier y sebagai fungsi dari x.

Bagi dataset kedalam data training dan data testing

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

Lakukan liner regresi sederhana pada data training

```
lm.r <- lm(formula = MonthsEmployed ~ Age,
           data = trainingset)
coef(lm.r)
```

```
## (Intercept)         Age
##   -2.354310    0.995091
```

Berdasarkan output diatas bisa dijelaskan bahwa: Persamaan regresi linernya adalah MonthsEmployed = 5,91 + 0,75*Age

Sebelum menggunakan rumus ini untuk memprediksi Bulan Bekerja di masa mendatang, Anda harus memastikan bahwa model ini signifikan secara statistik, yaitu: Ada hubungan yang signifikan secara statistik antara prediktor dan variabel hasil, kita akan memeriksa kualitas model regresi linier.

Summary dari model

```
summary(lm.r)
```

```
##
## Call:
## lm(formula = MonthsEmployed ~ Age, data = trainingset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.322 -21.180  -7.525  16.500  98.467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.3543     5.9568  -0.395    0.693
## Age           0.9951     0.1648   6.037  4.7e-09 ***
```

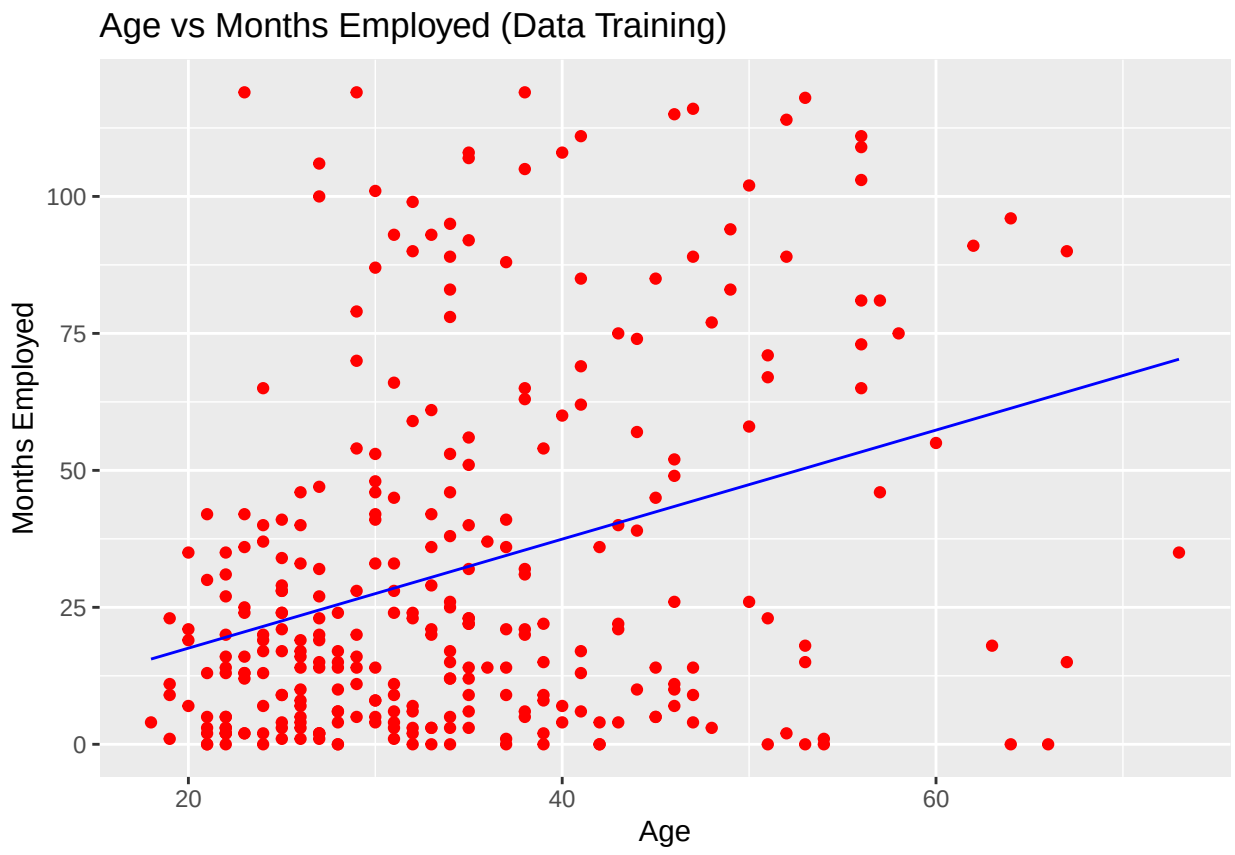
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.92 on 296 degrees of freedom
## Multiple R-squared:  0.1096, Adjusted R-squared:  0.1066
## F-statistic: 36.44 on 1 and 296 DF,  p-value: 4.704e-09
```

Prediksi model regresi dengan data testing

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

Visualasi hasil data training

```
ggplot() + geom_point(aes(x = trainingset$Age,
                          y = trainingset$MonthsEmployed), colour = 'red') +
  geom_line(aes(x = trainingset$Age,
                y = predict(lm.r, newdata = trainingset)), colour = 'blue') +
  ggtitle('Age vs Months Employed (Data Training)') +
  xlab('Age') +
  ylab('Months Employed')
```



Visualasi hasil data testing

```
ggplot() + geom_point(aes(x = testingset$Age,
                          y = testingset$MonthsEmployed), colour = 'red') +
  geom_line(aes(x = testingset$Age,
                y = predict(lm.r, newdata = testingset)), colour = 'blue') +
  ggtitle('Age vs Months Employed (Data Testing)') +
  xlab('Age') +
  ylab('Months Employed')
```