Modul 08 - Simple Linear Regression

Roni Yunis

12/04/2023

Pengantar

Regresi linear sederhana adalah suatu metode statistik yang digunakan untuk memodelkan hubungan linier antara satu variabel bebas (independen) dengan satu variabel terikat (dependen). Dalam konteks ini, "linier" mengacu pada hubungan garis lurus antara variabel-variabel tersebut.

Dalam persamaan regresi linear sederhana, dapat dinyatakan sebagai berikut:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y adalah variabel terikat (dependen).
- X adalah variabel bebas (independen).
- $\beta 0$ adalah intercept (konstanta).
- $\beta 1$ adalah koefisien regresi yang mengukur tingkat perubahan dalam Y untuk setiap satu unit perubahan dalam X.
- ε adalah kesalahan acak yang tidak dapat dijelaskan oleh model dan diasumsikan mengikuti distribusi normal dengan mean nol.

Tujuan utama dari regresi linear sederhana adalah menemukan nilai-nilai $\beta 0$ dan $\beta 1$ yang meminimalkan jumlah kuadrat kesalahan (sum of squared errors) antara nilai prediksi yang diberikan oleh model dan nilai aktual dari variabel terikat.

Regresi linear sederhana dapat digunakan untuk memahami dan memodelkan hubungan antara dua variabel, serta melakukan prediksi berdasarkan data yang ada. Metode ini umum digunakan dalam statistika dan analisis data.

Contoh:

Anggaplah Anda bekerja dalam sebuah perusahaan ritel dan Anda ingin memahami hubungan antara jumlah uang yang dihabiskan oleh pelanggan dalam satu transaksi X dengan total pendapatan penjualan Y pada transaksi tersebut. Anda dapat menggunakan regresi linear sederhana untuk memodelkan hubungan tersebut.

Contoh Persamaan Regresi Linear Sederhana:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y: Total pendapatan penjualan dalam satu transaksi.
- X: Jumlah uang yang dihabiskan oleh pelanggan dalam satu transaksi.
- β_0: Intercept (konstanta), mewakili pendapatan penjualan yang dihasilkan ketika X sama dengan nol.
- β_1: Koefisien regresi, menunjukkan seberapa banyak pendapatan penjualan meningkat untuk setiap peningkatan satu unit dalam X.

- ε: Kesalahan acak.

Anda mengumpulkan data dari sejumlah transaksi dan kemudian menggunakan regresi linear sederhana untuk mengestimasi nilai β_0 dan β_1. Setelah memperoleh model regresi, Anda dapat menggunakannya untuk membuat prediksi. Misalnya, jika seorang pelanggan menghabiskan $50 dalam transaksi, berapa total pendapatan penjualan yang diperkirakan?

Contoh Implementasi dengan R

Load Packages

```
#Split dataset
library (caTools)
```

```
## Warning: package 'caTools' was built under R version 4.3.2
```

```
#Predicting result visualization
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
#Import dataset marketing pada library datarium
library(datarium)
```

```
## Warning: package 'datarium' was built under R version 4.3.2
```

```
#library manipulasi data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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 glimpse()

```
glimpse(marketing)
```

```
## Rows: 200
## Columns: 4
## $ youtube   <dbl> 276.12, 53.40, 20.64, 181.80, 216.96, 10.44, 69.00, 144.24, ~
## $ facebook  <dbl> 45.36, 47.16, 55.08, 49.56, 12.96, 58.68, 39.36, 23.52, 2.52~
## $ newspaper <dbl> 83.04, 54.12, 83.16, 70.20, 70.08, 90.00, 28.20, 13.92, 1.20~
## $ sales     <dbl> 26.52, 12.48, 11.16, 22.20, 15.48, 8.64, 14.16, 15.84, 5.76,~
```

Exploratory Data Analysis

```
# Melihat summary data
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
```

```
# Melihat korelasi atau hubungan antar variabel
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

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

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

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Y = Sales X = Facebook

$$Sales = \beta_0 + \beta_1 facebook + \varepsilon$$

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

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

```r
dim(trainingset)
```

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

```r
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

```r
# Analisis regresi untuk variabel facebook terhap sales

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

```
##
## Call:
## lm(formula = sales ~ facebook, data = trainingset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.904  -2.744   1.012   3.400   9.247
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.90568    0.86659  12.585  < 2e-16 ***
## facebook     0.20872    0.02651   7.875 9.08e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.388 on 138 degrees of freedom
## Multiple R-squared:   0.31,  Adjusted R-squared:  0.305
## F-statistic: 62.01 on 1 and 138 DF,  p-value: 9.08e-13
```

$$Sales = 10.93519 + 0.21232.facebook + \varepsilon$$

Misalnya facebooknya Rp. 3500, berapakah nilai penjualan (sales)? Persamaan regresi linernya adalah:

$$Sales = 10.93519 + 0.21232 * 3500 + \varepsilon$$

```
sales = 10.93519 + (0.21232*3500)
sales
```

```
## [1] 754.0552
```

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

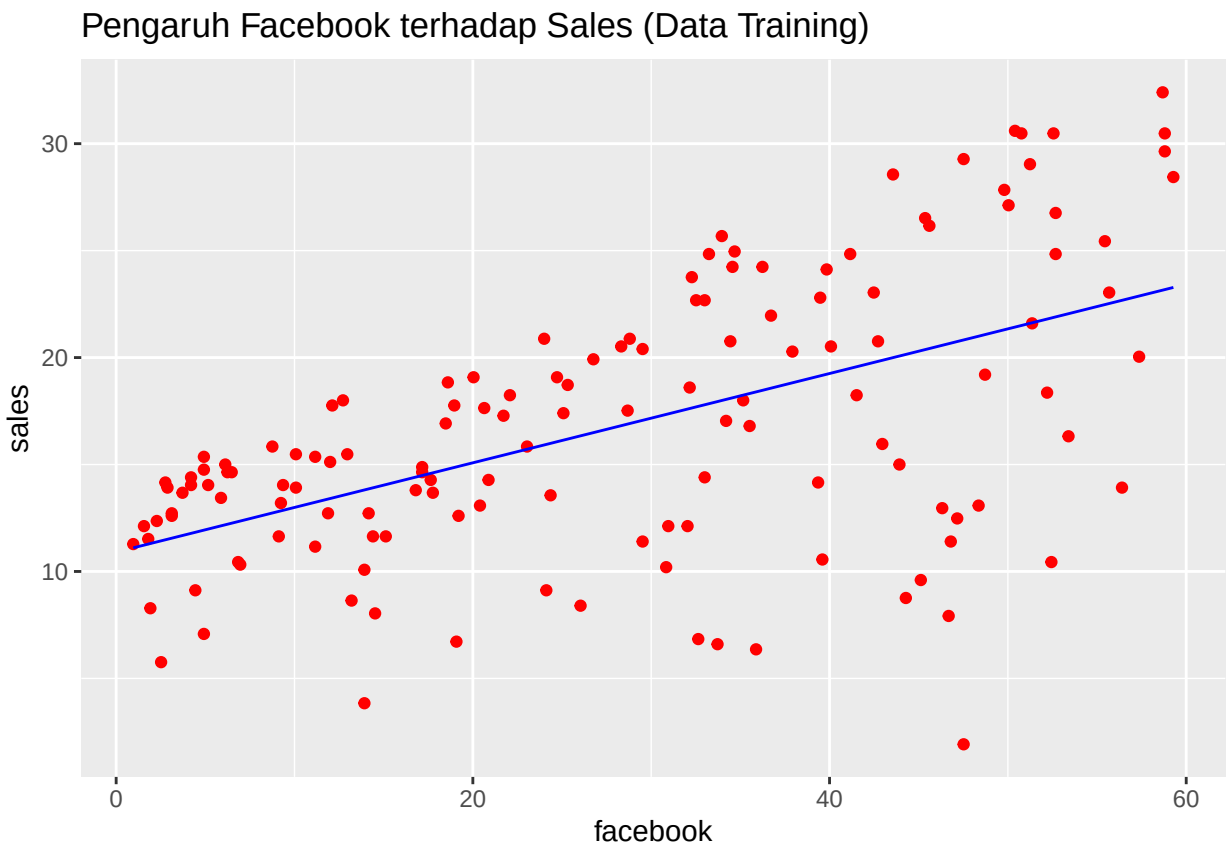
Prediksi model regresi dengan data testing

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

```
##        3        4        6        8       13       16
## 22.40182 21.24970 23.15320 15.81471 19.69684 22.85265
```

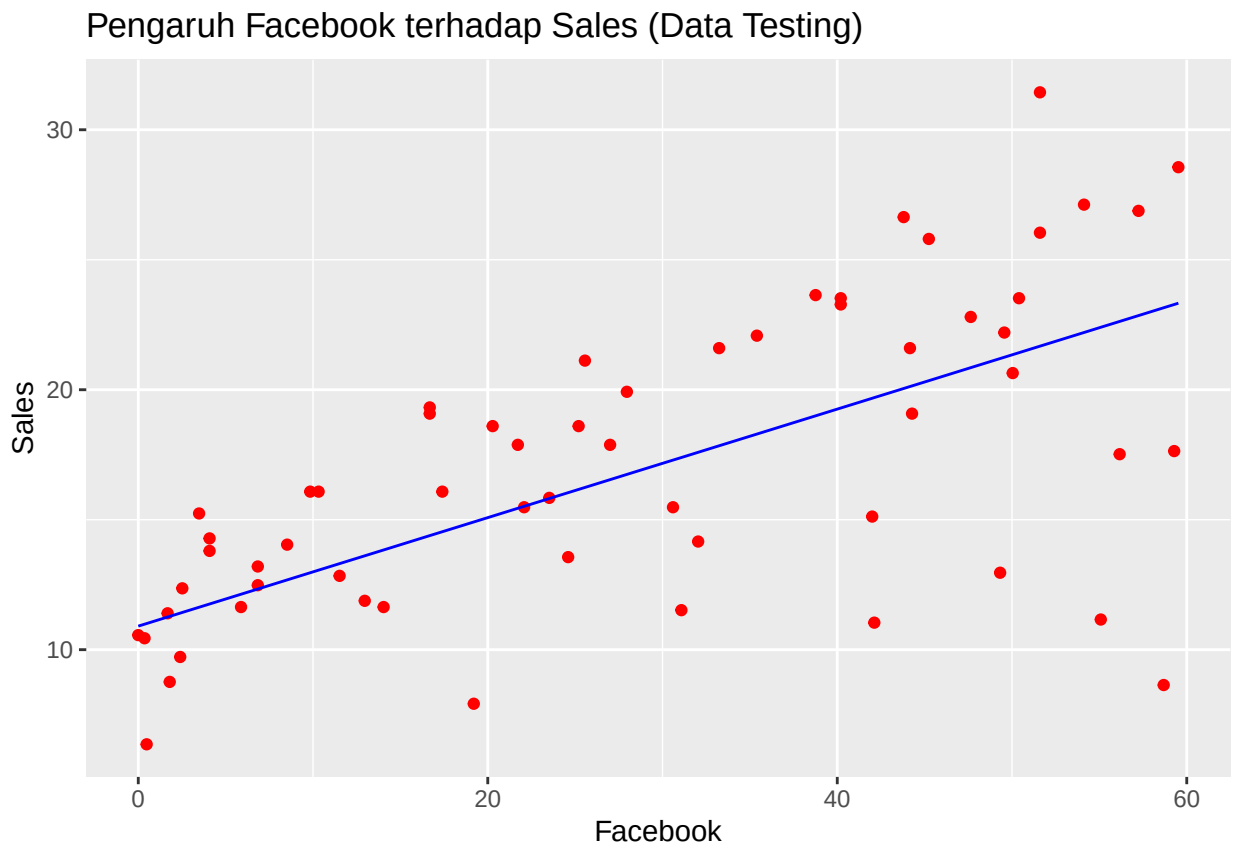
Visualasi hasil data training

```
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')
```



Evaluasi

Sum of Squared Errors (SSE) adalah metrik evaluasi yang mengukur jumlah kuadrat dari selisih antara nilai prediksi dari model regresi dan nilai aktual dalam data. SSE dapat memberikan gambaran tentang seberapa baik model sesuai dengan data.

Dalam konteks regresi linear sederhana, SSE dihitung dengan menjumlahkan kuadrat dari selisih antara nilai prediksi $\hat{Y}_i$ dan nilai aktual Y_i untuk setiap observasi:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Dalam R, Anda dapat mengakses SSE langsung dari objek model regresi menggunakan fungsi `sum(model$residuals^2)`. Semakin kecil nilai SSE, semakin baik modelnya, karena SSE mencerminkan seberapa baik model dapat memperkirakan nilai aktual dalam dataset. Namun, penggunaan SSE sebaiknya selalu dikombinasikan dengan metrik evaluasi lainnya, seperti R-squared, MSE, atau uji lainnya, untuk mendapatkan gambaran yang lengkap tentang kinerja model.

```
# Menghitung SSE
sse <- sum(lm.r$residuals^2)
cat("Sum of Squared Errors (SSE):", sse, "\n")
```

```
## Sum of Squared Errors (SSE): 4006.082
```

R-Squared, Nilai R-squared mengindikasikan seberapa besar variasi dalam variabel terikat yang dapat dijelaskan oleh model. Nilai R-squared berkisar antara 0 (model tidak menjelaskan variasi sama sekali) hingga 1 (model menjelaskan seluruh variasi). Anda dapat menghitung R-squared (koefisien determinasi) dari model regresi linear sederhana menggunakan fungsi `summary()` pada objek model

```
# Mengitung R-Square
model.summary <- summary(lm.r)
r_squared <- model.summary$r.squared
cat("R-squared:", r_squared, "\n")
```

```
## R-squared: 0.3100318
```

Mean Squared Error (MSE) Menghitung MSE dapat memberikan gambaran tentang seberapa baik model dapat menjelaskan variasi dalam data. Semakin rendah MSE dan semakin tinggi R^2, semakin baik modelnya.

```
# Menghitung MSE
mse <- mean(lm.r$residuals^2)
cat("Mean Squared Error (MSE):", mse, "\n")
```

```
## Mean Squared Error (MSE): 28.61487
```

Metode evaluasi ini memberikan pandangan holistik tentang kinerja model regresi linear sederhana, dan pemilihan metode yang sesuai tergantung pada kebutuhan spesifik analisis Anda.

Latihan

1. Buatlah model regresi untuk hubungan variabel `youtube` terhadap variabel `sales`
2. Buatlah model regresi unutk hubungan variabel `newpaper` terhadap variabel `sales`

1. Model Regresi Youtube terhadap Sales

```
# menghitung korelasi antar variabel
koryoutube <- cor(marketing$youtube, marketing$sales)
koryoutube
```

```
## [1] 0.7822244
```

Nilai korelasi dari kedua variabel adalah 0.7822244

Persamaan regresi dari youtube terhadap sales:

$$Sales = \beta_0 + \beta_1 youtube + \varepsilon$$

```
# liner regresi sederhana pada data training

lm.y <- lm(sales ~ youtube,
           data = trainingset)
summary(lm.y)
```

```
##
## Call:
## lm(formula = sales ~ youtube, data = trainingset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0589  -2.1218  -0.1318   2.6290   8.7261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.196705   0.645423   12.70   <2e-16 ***
## youtube     0.048254   0.003135   15.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.935 on 138 degrees of freedom
## Multiple R-squared:  0.632,  Adjusted R-squared:  0.6293
## F-statistic:   237 on 1 and 138 DF,  p-value: < 2.2e-16
```

Berdasarkan output diatas bisa dijelaskan bahwa: Persamaan regresi linernya adalah

$$sales = 8.175714 + 0.050387 * youtube$$

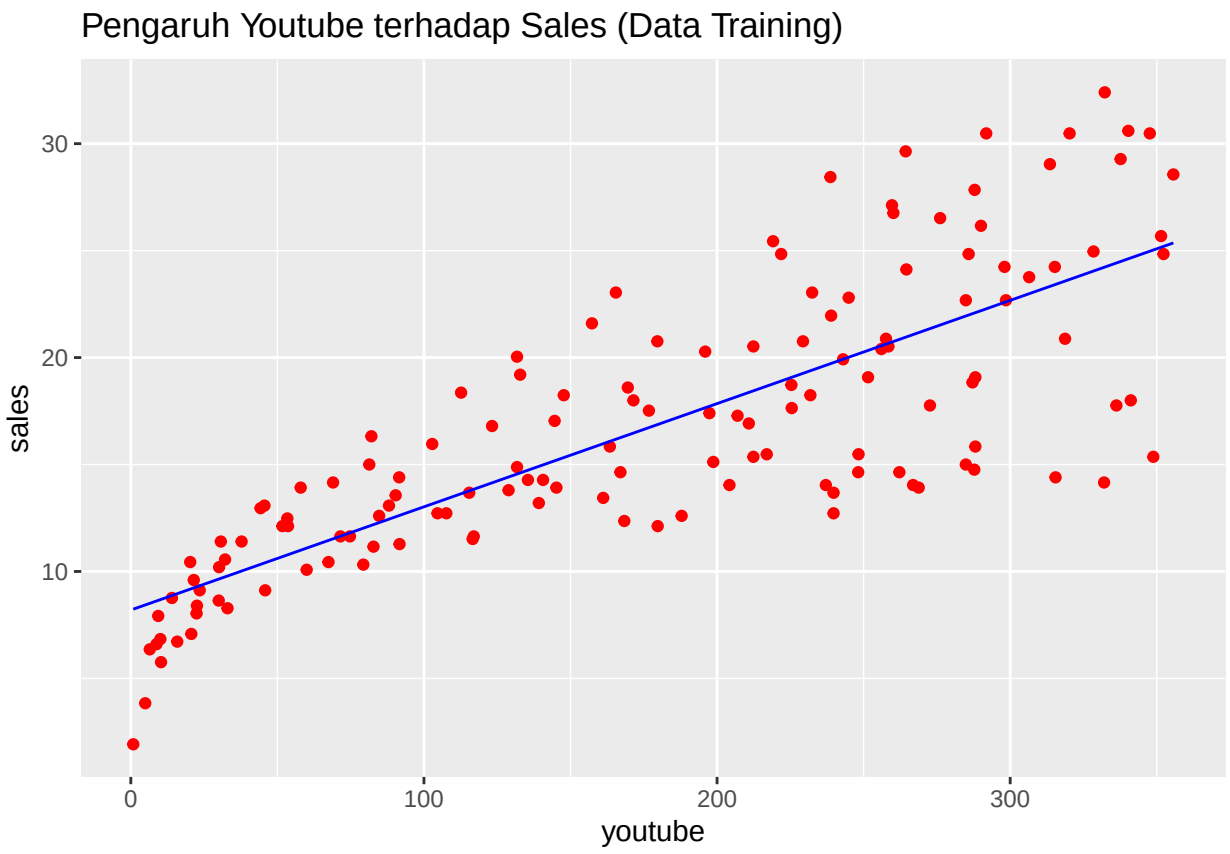
Misal youtube = 3500

```
sales.y = 8.175714 + 0.050387*3500
sales.y
```

```
## [1] 184.5302
```

Visualasi hasil data training youtube

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



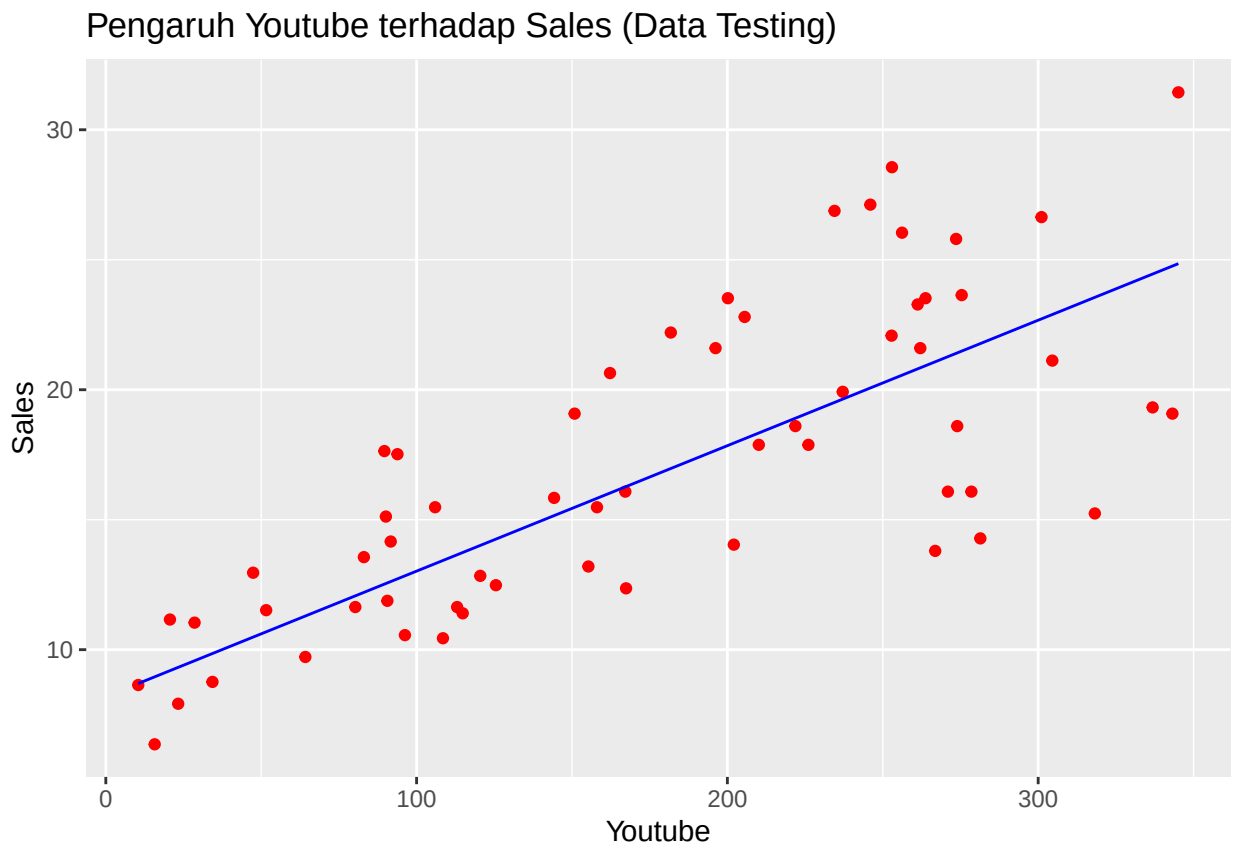
Prediksi model regresi dengan data testing

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

```
##         3         4         6         8        13        16
##  9.192666 16.969266  8.700476 15.156849  9.574837 19.511282
```

Visualasi hasil data testing

```
ggplot() + geom_point(aes(x = testingset$youtube,
                          y = testingset$sales), colour = 'red') +
  geom_line(aes(x = testingset$youtube,
                y = predict(lm.y, newdata = testingset)), colour = 'blue') +
  ggtitle('Pengaruh Youtube terhadap Sales (Data Testing)') +
  xlab('Youtube') +
  ylab('Sales')
```



Latihan ## Evaluasi Lakukan evaluasi dari model yang sudah dibuat dengan menggunakan SSE, MSE dan R-Squared

```
# Evaluasi SSE youtube terhadap sales

sse_y <- sum(lm.y$residuals^2)
cat("Sum of Squared Errors (SSE):", sse_y, "\n")
```

```
## Sum of Squared Errors (SSE): 2136.843
```

```
# Evaluasi RSquare youtube terhadap sales
model.summary_y <- summary(lm.y)
r_squared_y <- model.summary_y$r.squared
cat("R-squared:", r_squared_y, "\n")
```

```
## R-squared: 0.6319712
```

```
# Evaluasi MSE youtube terhadap sales
mse_y <- mean(lm.y$residuals^2)
cat("Mean Squared Error (MSE):", mse_y, "\n")
```

```
## Mean Squared Error (MSE): 15.26316
```

2. Model Regresi Newspaper terhadap Sales

Buatlah model regresi untuk hubungan varibel `newspaper` terhadap variabel `sales`

```
# Melihar korelasi newspaper dengan sales

kornewspaper <- cor(marketing$newspaper, marketing$sales)
kornewspaper
```

```
## [1] 0.228299
```

Persamaan regresi liner antara `newspaper` dengan `sales`

$$Sales = \beta_0 + \beta_1.newspaper + \varepsilon$$

```
# Model regresi liner newspaper terhadap sales
lm.n <- lm(sales ~ newspaper,
           data = trainingset)
summary(lm.n)
```

```
##
## Call:
## lm(formula = sales ~ newspaper, data = trainingset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.402  -4.088  -1.068   4.373  15.410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.73011    0.90669  16.246  < 2e-16 ***
## newspaper    0.05667    0.02095   2.705  0.00769 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.321 on 138 degrees of freedom
## Multiple R-squared:  0.05035,    Adjusted R-squared:  0.04347
## F-statistic: 7.317 on 1 and 138 DF,  p-value: 0.007693
```

$$Sales = 15.09358 + 0.05201 * newspaper + \varepsilon$$

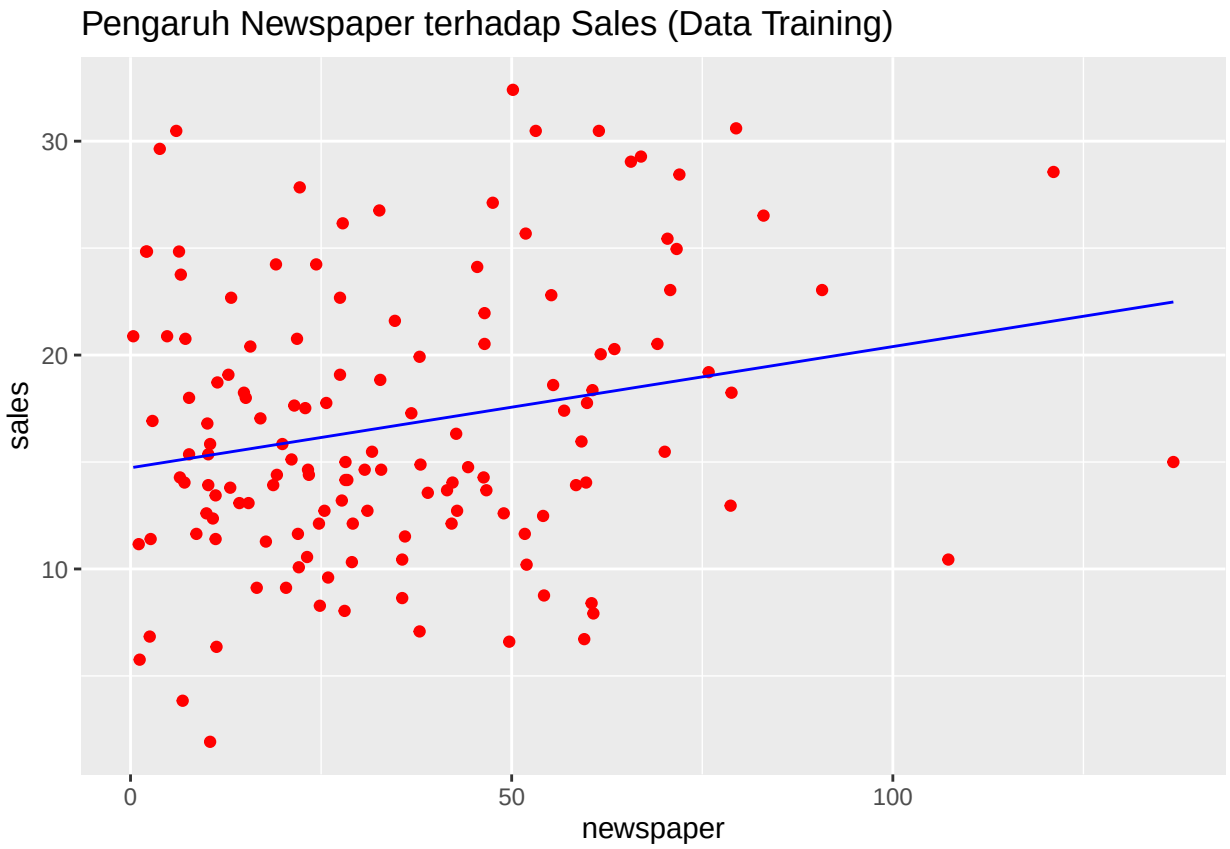
Misal nilai newspapernya: 3500 maka nilai:

```
sales_n = 15.09358 + (0.05201 * 3500)
sales_n
```

```
## [1] 197.1286
```

Visualasi hasil data training newspaper

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



Prediksi model dengan data testing

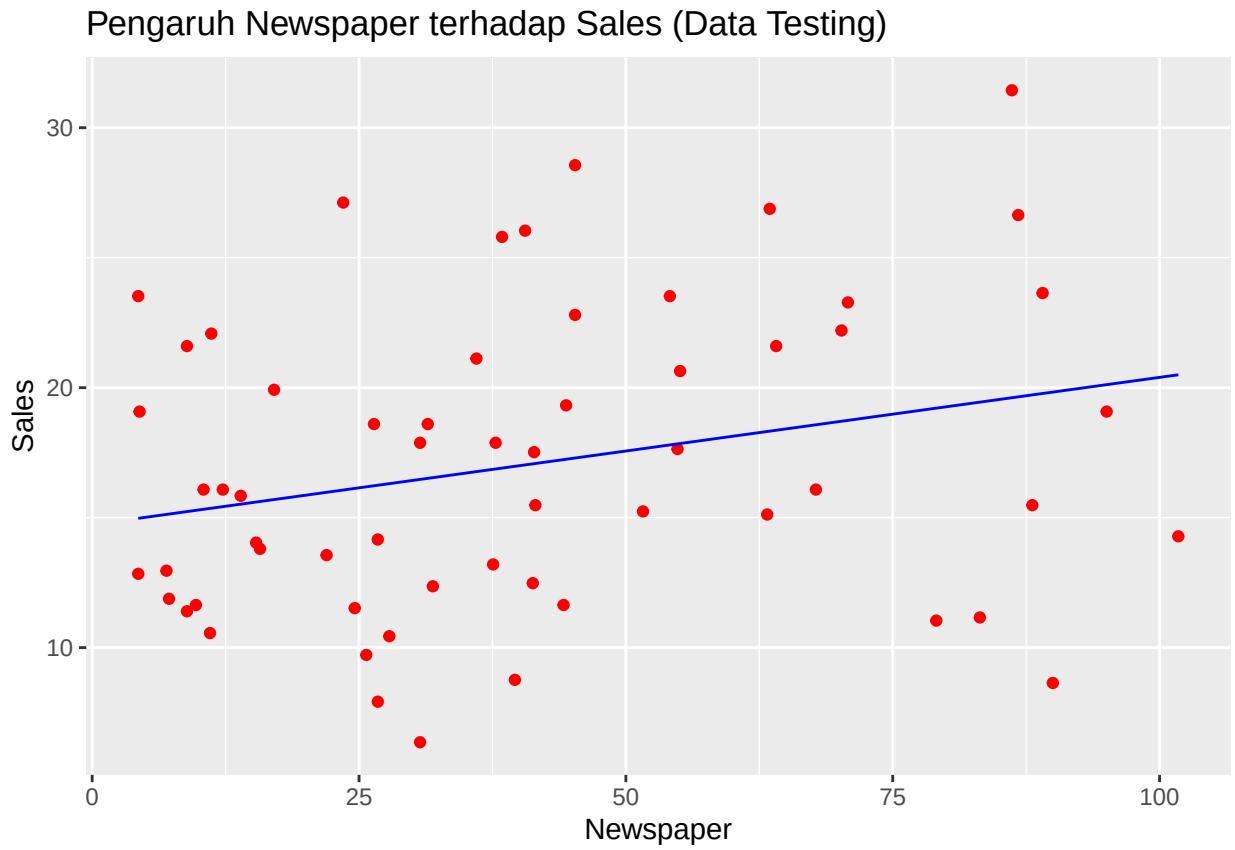
```
prediksi.n <- predict(lm.n, newdata = testingset)
head(prediksi.n)
```

```
##        3        4        6        8       13       16
## 19.44271 18.70828 19.83032 15.51894 19.21150 18.32746
```

Visualasi hasil data testing

```
ggplot() + geom_point(aes(x = testingset$newspaper,
                          y = testingset$sales), colour = 'red') +
  geom_line(aes(x = testingset$newspaper,
                y = predict(lm.n, newdata = testingset)), colour = 'blue') +
  ggtitle('Pengaruh Newspaper terhadap Sales (Data Testing)') +
```

```
  xlab('Newspaper') +
  ylab('Sales')
```



Evaluasi model

```
# Evaluasi SSE newspaper  terhadap sales

sse_n <- sum(lm.n$residuals^2)
cat("Sum of Squared Errors (SSE)", sse_n, "\n")
```

```
## Sum of Squared Errors (SSE) 5513.845
```

```
# Evalusasi RSquare newpaper terhadap sales
model.summary_n <- summary(lm.n)
r_squared_n <- model.summary_n$r.squared
cat("R-squared:", r_squared_n, "\n")
```

```
## R-squared: 0.05034953
```

```
# Evaluasi MSE newspaper terhadap sales
mse_n <- mean(lm.n$residuals^2)
cat("Mean Squared Error (MSE):", mse_n, "\n")
```

```
## Mean Squared Error (MSE): 39.38461
```

```r
model_performance <- data.frame(
  No = c(1:3),
  Variable = c("FACEBOOK", "YOUTUBE", "NEWSPAPER"),
  SSE = c(sse, sse_y, sse_n),
  Rquare = c(r_squared, r_squared_y, r_squared_n),
  MSE = c(mse, mse_y, mse_n),
  stringsAsFactors = FALSE

)

model_performance
```

```
##   No  Variable      SSE     Rquare      MSE
## 1  1  FACEBOOK 4006.082 0.31003183 28.61487
## 2  2   YOUTUBE 2136.843 0.63197125 15.26316
## 3  3 NEWSPAPER 5513.845 0.05034953 39.38461
```

Kesimpulan: Berdasarkan 3 variabel prediktor yaitu facebook, youtube, dan newspaper yang paling berpengaruh terhadap peningkatakn penjualan adalah variabel youtube. Hal ini bisa dijelaskan dari nilai evaluasi SSE dan MSE bahwa model youtube dan sales lebih kecil dan RSquare lebih besar dibandingkan dengan model regresi dari 2 variabel lainnya.