

Modul 09 - Multiple Linear Regression

Roni Yunis

12/09/2024

Pengantar

Regresi Liner Berganda adalah bentuk lain dari regresi linier sederhana yang digunakan untuk memprediksi variabel Y (dependent), berdasarkan beberapa variabel prediktor X (independent) Kalau dengan 3 variabel prediktor, maka prediksi Y bisa dinyatakan dalam persamaan berikut:

$$y = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \beta_3.x_3 + \epsilon$$

Nilai β disebut dengan bobot regresi (koefisien beta), digunakan untuk mengukur hubungan antara variabel prediktor dan hasil. β_j dapat diartikan sebagai efek rata-rata pada y dari peningkatan satu unit dalam x_j , dimana semua prediktor lainnya tetap. Dalam Modul 09 ini kita akan membahas:

1. Bagaimana membangun model regresi berganda dan bagaimana cara menginterpretasikannya
2. Memeriksa kualitas dari model yang sudah dihasilkan

Data Preparation

Dalam kasus ini, kita akan menggunakan dataset yang ada pada packages datarium, nama datasetnya adalah **marketing**, sebelum kita menggunakan dataset tersebut, kita akan install dulu **library (datarium)**

```
#Manipulasi data  
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
##
```

```
## 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
```

```
#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 4.3.3
```

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

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

```
#menampilkan isi dataset 6 baris teratas  
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
```

Kita akan melihat ringkasan data dari dataset marketing

```
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,~
```

```
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
```

Bisa dilihat bahwa, ada 4 buah kolom youtube, facebook, newspaper, dan sales

```
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
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
# Membuat korelasi dengan matrik
```

```
marketing_cor <- cor(marketing)
```

```
corrplot(marketing_cor, method = "color", addCoef.col = "green")
```



```
library(GGally)
```

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

```
## Registered S3 method overwritten by 'GGally':
```

```
##   method from
```

```
##   +.gg      ggplot2
```

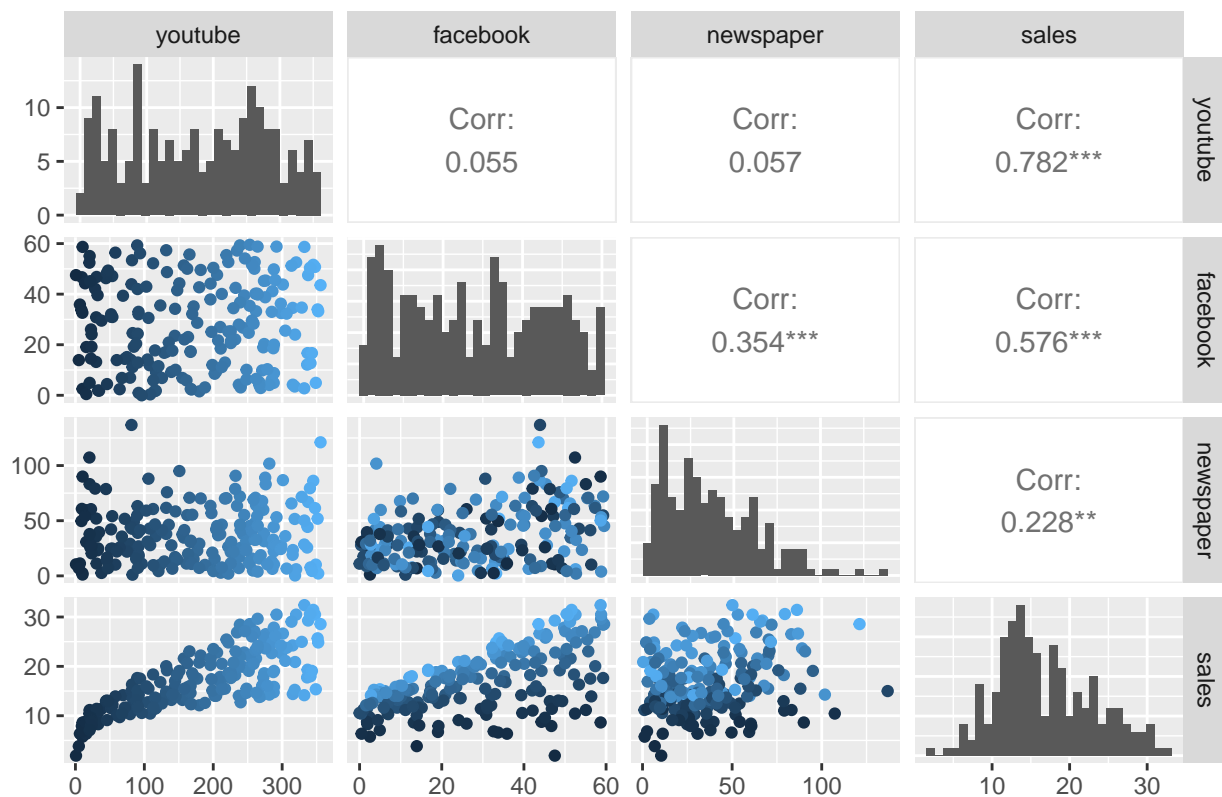
```
# Buat correlogram
ggpairs(marketing,
  columns = 1:4, # Pilih kolom yang akan diplot
  title = "Correlation Multiple Variables",
  lower = list(continuous = "points", mapping = aes(color = youtube)),
  diag = list(continuous = "barDiag"))
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

Correlation Multiple Variables

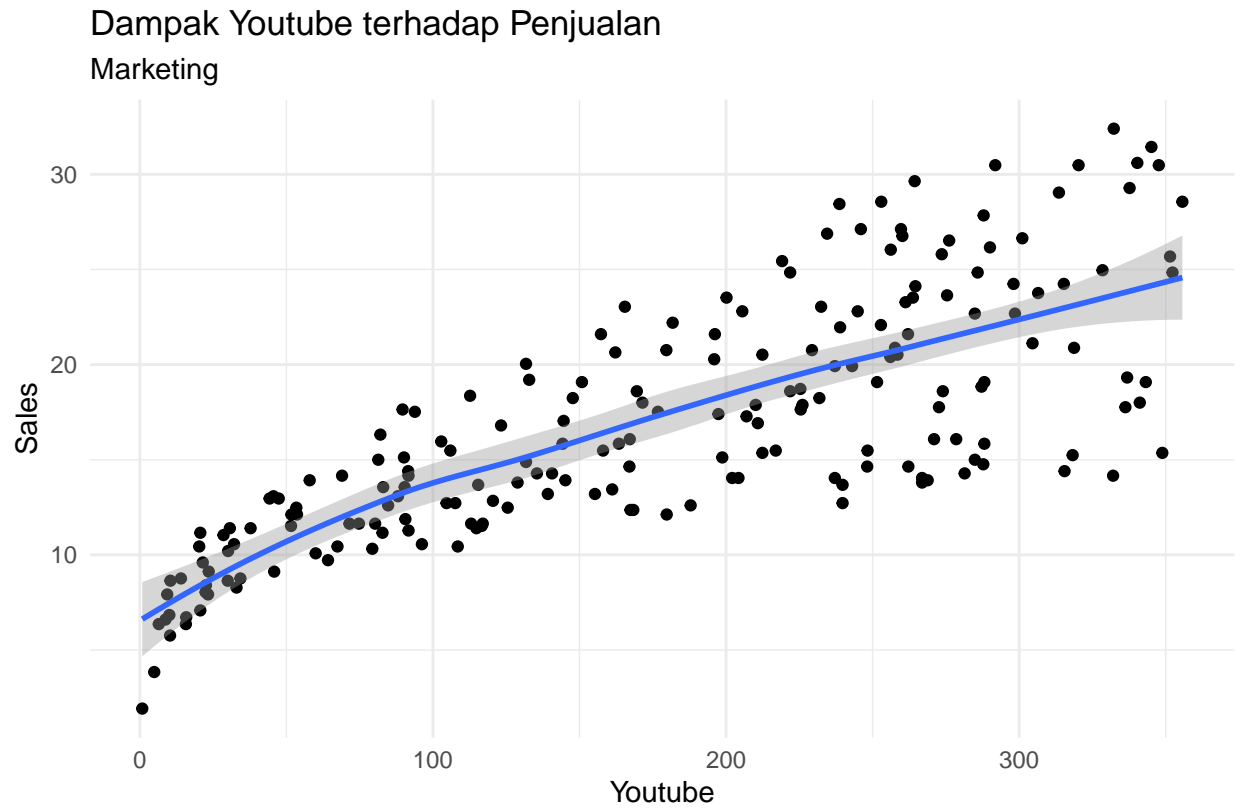


```
# Visualisasi dampak youtube pada sales
#library(ggplot2)

ggplot(marketing,
  aes(youtube, sales)) +
  geom_point() +
  geom_smooth() +
  labs(
    title = "Dampak Youtube terhadap Penjualan",
    subtitle = "Marketing",
    caption = "by: Roni Yunis",
```

```
x = "Youtube",
y = "Sales"
) +
theme_minimal()
```

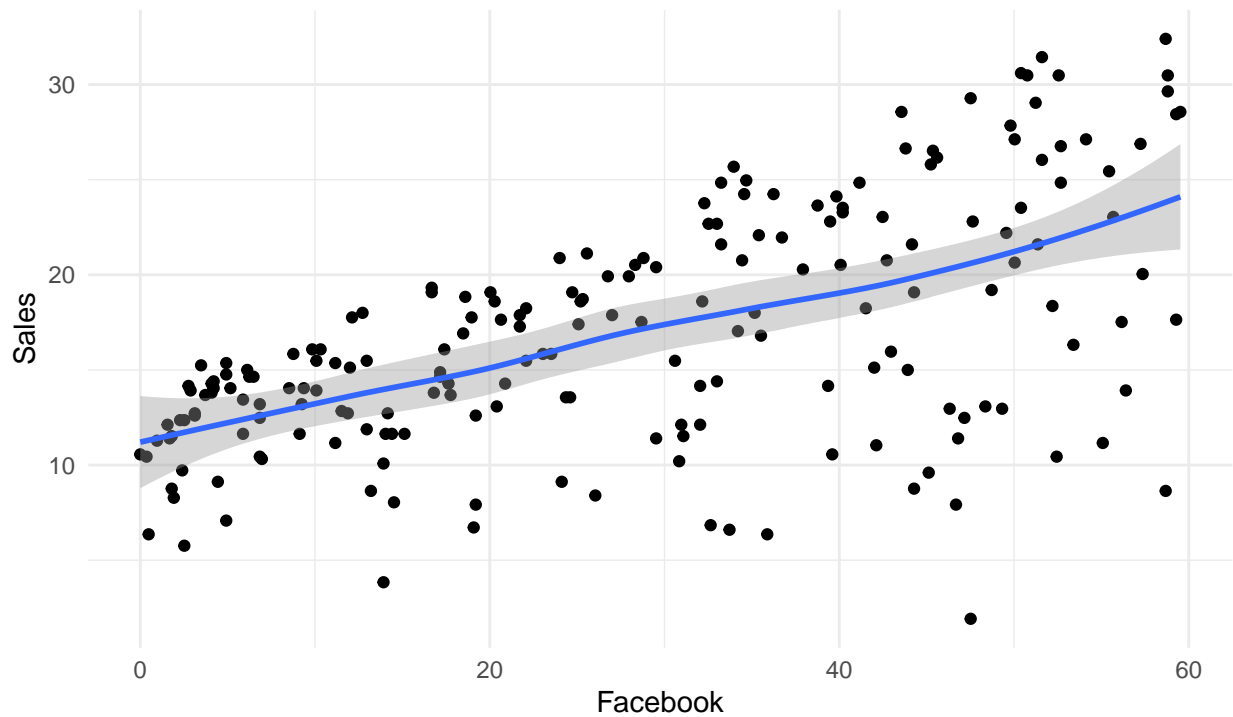
```
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
```



```
# Visualisasi dampak facebook pada sales
library(ggplot2)
ggplot(marketing,
       aes facebook, sales)) +
  geom_point() +
  geom_smooth() +
  labs(
    title = "Dampak Facebook terhadap Penjualan",
    subtitle = "Marketing",
    caption = "by: Roni Yunis",
    x = "Facebook",
    y = "Sales"
  ) +
  theme_minimal()
```

```
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
```

Dampak Facebook terhadap Penjualan Marketing

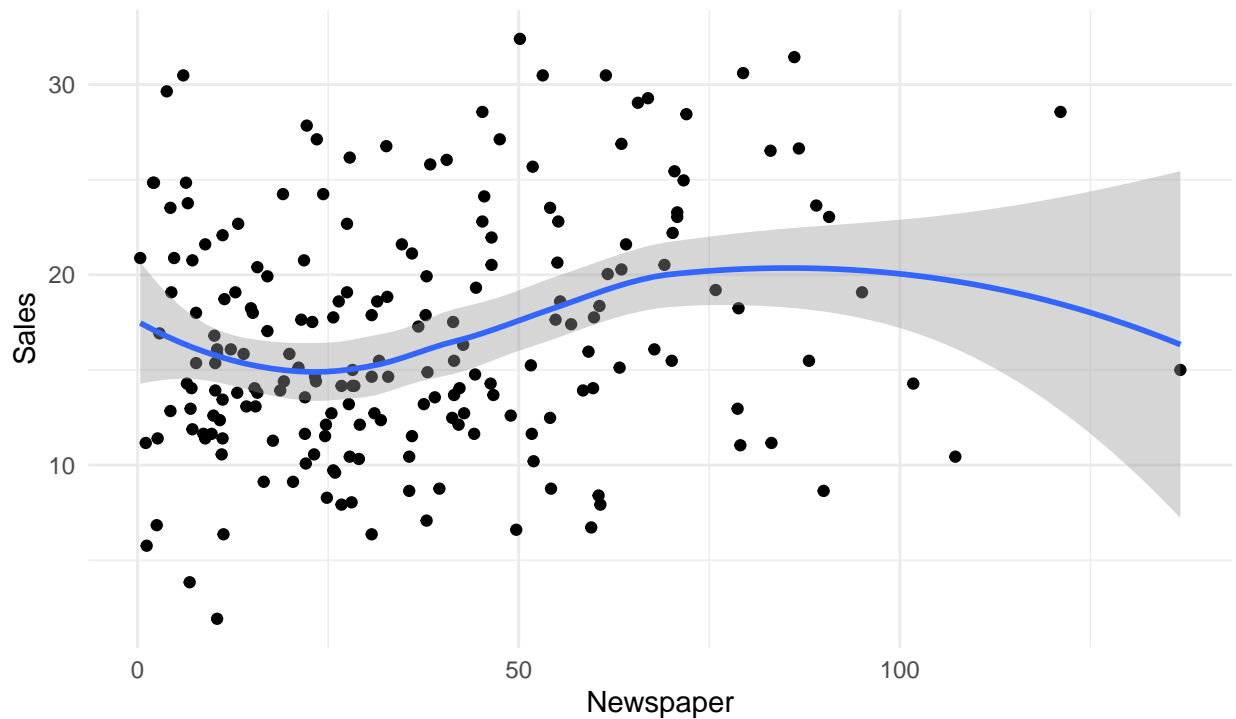


by: Roni Yunis

```
# Visualisasi dampak Newspaper pada sales
library(ggplot2)
ggplot(marketing,
       aes(newspaper, sales)) +
  geom_point() +
  geom_smooth() +
  labs(
    title = "Dampak Newspaper terhadap Penjualan",
    subtitle = "Marketing",
    caption = "by: Roni Yunis",
    x = "Newspaper",
    y = "Sales"
  ) +
  theme_minimal()
```

```
## 'geom_smooth()' using method = 'loess' and formula = 'y ~ x'
```

Dampak Newspaper terhadap Penjualan Marketing



by: Roni Yunis

Model Regresi dan Interpretasi

Sekarang kita akan membuat model regresi untuk memprediksi tingkat penjualan dari biaya iklan yang sudah dimuat pada youtube, facebook dan newspaper. Kalau kita buat model regresinya, maka sales adalah variabel dependen (y), youtube, facebook, dan newspaper adalah variabel independen (x). Sehingga model regresinya bisa didefinisikan menjadi seperti persamaan berikut:

$$sales = \beta_0 + \beta_1.youtube + \beta_2.facebook + \beta_3.newspaper$$

Model regresinya akan kita simpan kedalam objek *liner*

```
liner_1 <- lm(sales ~ youtube + facebook + newspaper, data = marketing)
summary(liner_1)
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5932  -1.0690   0.2902   1.4272   3.3951
##
## Coefficients:
```

```
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.526667  0.374290  9.422  <2e-16 ***
## youtube      0.045765  0.001395 32.809  <2e-16 ***
## facebook     0.188530  0.008611 21.893  <2e-16 ***
## newspaper    -0.001037  0.005871 -0.177    0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

p-value youtube terhadap sales = 0.00000000000000022 < 0.05 (signifikan) p-value facebook terhadap sales = 0.00000000000000022 < 0.05 (signifikan)

Langkah pertama yang dapat kita gunakan untuk menafsirkan analisis regresi berganda yang sudah kita lakukan, adalah memeriksa nilai statistik F dan nilai p yang terkait, hal ini bisa kita lihat pada baris terakhir dari hasil model regresi. Dalam contoh kali ini, didapat bahwa nilai p-value dari F-Statistik adalah < 2.2e-16 (0.00000000000000022), artinya nilai ini adalah sangat signifikan. Jadi bisa disimpulkan bahwa salah satu dari variabel prediktor (independen) berhubungan secara signifikan dengan variabel hasil (dependen).

Untuk melihat variabel prediktor mana yang paling signifikan, kita dapat memeriksa nilai koefisiennya, tabel koefisien digunakan untuk melihat estimasi koefisien beta regresi dan nilai t-statistik p-value yang terkait. Untuk menghitung nilai koefisien regresi tsb, maka bisa kita tulis seperti ini:

```
summary(liner_1)$coefficient
```

```
##           Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)  3.526667243 0.374289884  9.4222884 1.267295e-17
## youtube      0.045764645 0.001394897 32.8086244 1.509960e-81
## facebook     0.188530017 0.008611234 21.8934961 1.505339e-54
## newspaper    -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

Kalau kita lihat dari hasil diatas (t value), terlihat bahwa anggaran iklan melalui youtube dan facebook berhubungan secara signifikan dengan perubahan penjualan. Tetapi anggaran untuk surat kabar tidak berhubungan signifikan. Sehingga bisa kita simpulkan bahwa jika anggaran \$1000 untuk iklan di *facebook* maka akan menyebabkan peningkatan rata-rata penjualan sebesar $0,1885 \times 1000 = 189$ unit penjualan. Jika kita lihat dari koefisien *youtube*, maka rata-rata peningkatan penjualan sebesar $0,045 \times 1000 = 45$ unit penjualan.

Jadi dari ketiga variabel yang ada, hanya 2 variabel yang mempengaruhi tingkat penjualan. Karena variabel *newspaper* tidak signifikan maka kita bisa perbaiki model regresi sebelumnya dengan cara tidak memasukkan variabel *newspaper* kedalamnya.

```
liner_2 <- lm(sales ~ youtube + facebook, data = marketing)
summary(liner_2)
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5572  -1.0502   0.2906   1.4049   3.3994
```



```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.50532    0.35339   9.919  <2e-16 ***
## youtube      0.04575    0.00139  32.909  <2e-16 ***
## facebook     0.18799    0.00804  23.382  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.018 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Sehingga berdasarkan model tersebut, maka kita bisa defenisikan persamaan dari model tersebut menjadi:

$$sales = 3.505 + 0,046.youtube + 0,188.facebook$$

Nilai anggaran iklan di youtube: 276.12, dan iklan di facebook: 45.36

$$sales = 3.505 + (0,046 * 276.12) + (0,188 * 45.36)$$

```
sales = 3.505 + (0.046*276.12) + (0.188*45.36)
sales
```

```
## [1] 24.7342
```

Nilai Aktual = 26.52, nilai Prediksi = 24.732. Nilai prediksi dengan nilai aktual ada selisih.

Nilai confiden interval dari model tersebut, bisa kita hitung dengan fungsi `confint()`

```
confint(liner_2)
```

```
##              2.5 %      97.5 %
## (Intercept) 2.80841159 4.20222820
## youtube     0.04301292 0.04849671
## facebook    0.17213877 0.20384969
```

Model Akurasi

Untuk melihat akurasi dari model liner yang sudah dihasilkan secara statistik bisa melihat pada hasil *Adjusted R-squared*, dalam contoh kasus ini nilainya adalah 0,896 artinya 89,6% dari nilai penjualan dapat ditingkatkan oleh anggaran dari iklan *youtube* dan *facebook*.

Selanjutnya adalah bagaimana cara kita mengukur kesalahan prediksi atau **Residual Standard Error (RSE)** atau menggunakan fungsi `sigma()`. Dalam sebuah model regresi yang baik, semakin rendah nilai RSE, maka akan semakin akurat model regresi tersebut. Untuk menghitungnya kita bisa membagi nilai RSE dengan rata-rata variabel hasil.

```
sigma(liner_2)/mean(marketing$sales)
```

```
## [1] 0.1199045
```

Jadi bisa dilihat bahwa tingkat kesalahan (error rate) dari model yang sudah dihasilkan adalah sebesar 12 %. Bisa kita simpulkan tingkat akurasi dari model adalah $100\% - 12\% = 88\%$

Latihan Dari model regresi yang sudah kita bahas sebelumnya, kelihatan bahwa hanya 2 variabel yang berpengaruh pada penjualan. Variabel yang tidak berpengaruh adalah variabel *newspaper*, sekarang coba Anda buktikan kalau hanya ada satu variabel yaitu *newspaper* saja apakah benar-benar tidak berpengaruh pada nilai penjualan?

$$sales = b_0 + b_1 * newspaper$$

```
# pangaruh newspaper terhadap sales

liner_3 <- lm(sales ~ newspaper, data = marketing)
summary(liner_3)

##
## Call:
## lm(formula = sales ~ newspaper, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.473  -4.065  -1.007   4.207  15.330
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.82169    0.74570   19.88  < 2e-16 ***
## newspaper     0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.111 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

Interpretasi: variabel newspaper berpengaruh positif dan signifikan terhadap sales, tapi pengaruhnya tidak besar atau relatif kecil, karena nilai R-square = 0.05212

Jika nilai iklan di newspaper 1000, maka berapa nilai penjualan (sales)?

$$sales = 14.82169 + (0.05469 * newspaper)$$
$$sales = 14.82169 + (0.05469 * 1000)$$

```
sales_newspaper = 14.82169 + (0.05469*1000)
sales_newspaper

## [1] 69.51169

sales_yf = 3.505 + (0.046*1000) + (0.188*1000)
sales_yf

## [1] 237.505
```

Model Regresi dengan Model GLM (Generalized Linear Model)

Bagi dataset kedalam data training dan data testing

```
splitdata <- sample.split(marketing$sales, SplitRatio = 0.7) # 70% untuk data training, sisanya data te
trainingset <- subset(marketing, splitdata == TRUE)
testingset <- subset(marketing, splitdata == FALSE)
```

```
dim(trainingset)
```

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

```
dim(testingset)
```

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

Model Regresi

Fit Model 1

```
# library model GLM
library(glm2)

# model GLM untuk memprediksi variabel sales pada data training
model_glm <- glm(sales ~ youtube + facebook + newspaper, data = trainingset,
                 family = gaussian(link = "identity"), # distribusi probabilitas dari variabel respons
                 control = list(epsilon = 1e-8, # nilai batasan toleransi kesalahan (0.0000001)
                               maxit = 100)      # jumlah iterasi
                 )

# melihat hasil model prediksi
summary(model_glm)
```

```
##
## Call:
## glm(formula = sales ~ youtube + facebook + newspaper, family = gaussian(link = "identity"),
##      data = trainingset, control = list(epsilon = 1e-08, maxit = 100))
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.993025   0.472703   6.332 3.28e-09 ***
## youtube      0.048507   0.001722  28.170 < 2e-16 ***
## facebook     0.183034   0.011135  16.437 < 2e-16 ***
## newspaper    0.002454   0.007287   0.337  0.737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 4.359657)
##
```

```
##      Null deviance: 5507.04  on 139  degrees of freedom
## Residual deviance:  592.91  on 136  degrees of freedom
## AIC: 609.38
##
## Number of Fisher Scoring iterations: 2
```

Latihan Coba diidentifikasi jenis family secara umum dalam GLM?, kemudian coba implementasikan ke dalam model. Kalau bisa anda bandingkan hasilnya. Model dengan hyperparameter mana yang paling baik?

Prediksi data testing dengan model

```
# Melakukan prediksi sales dengan model GLM pada data testing
predicted_glm <- predict(model_glm, newdata = testingset, type = "response")

# Menampilkan hasil prediksi
head(predicted_glm)
```

```
##      2      5      10      13      14      16
## 14.34799 16.06113 15.25647 12.28185 10.35878 24.99953
```

```
head(testingset$sales)
```

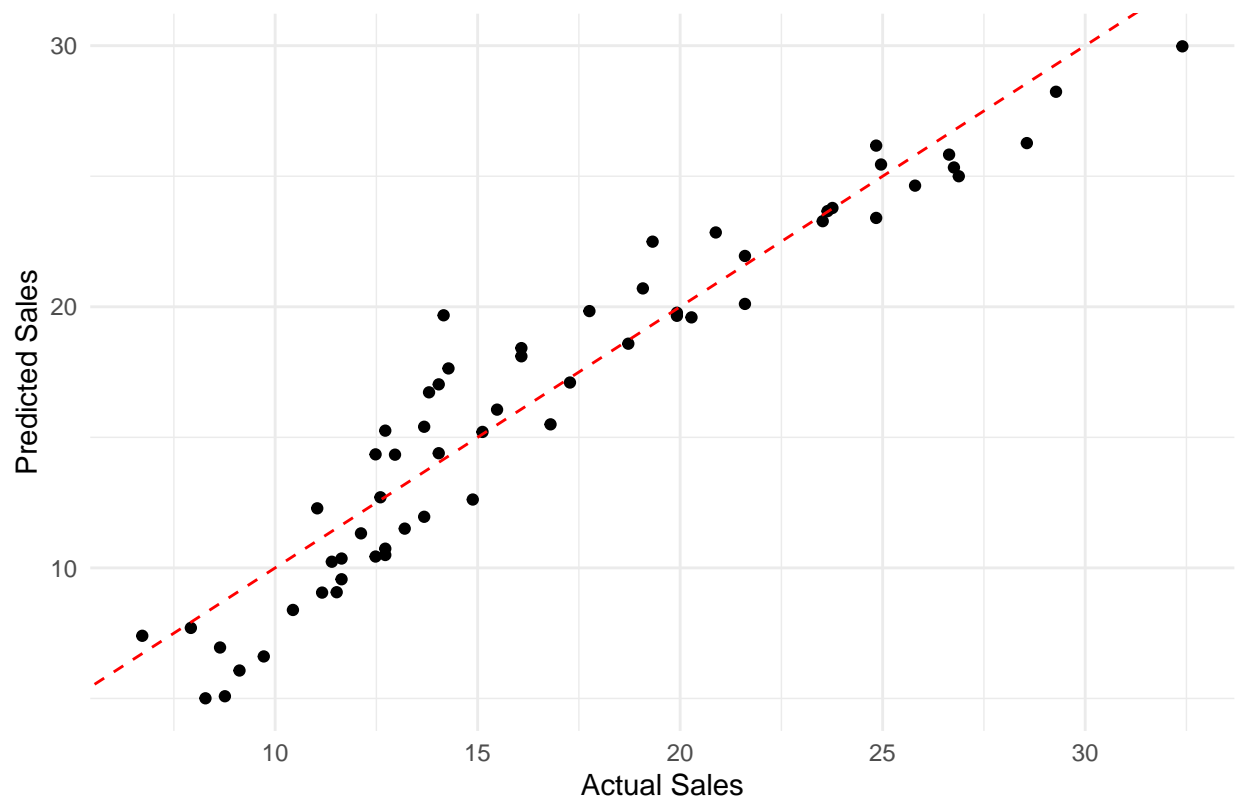
```
## [1] 12.48 15.48 12.72 11.04 11.64 26.88
```

Visualisasi Hasil Prediksi

```
# Visualisasi Hasil
result_data_glm <- data.frame(sales = testingset$sales, Predictions = predicted_glm)

# Visualisasi Perbandingan Sales Aktual dengan Hasil Prediksi
ggplot(data = result_data_glm, aes(x = sales, y = Predictions)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  labs(x = "Actual Sales", y = "Predicted Sales") +
  ggtitle("Comparison of Actual Sales and GLM Model Prediction") +
  theme_minimal()
```

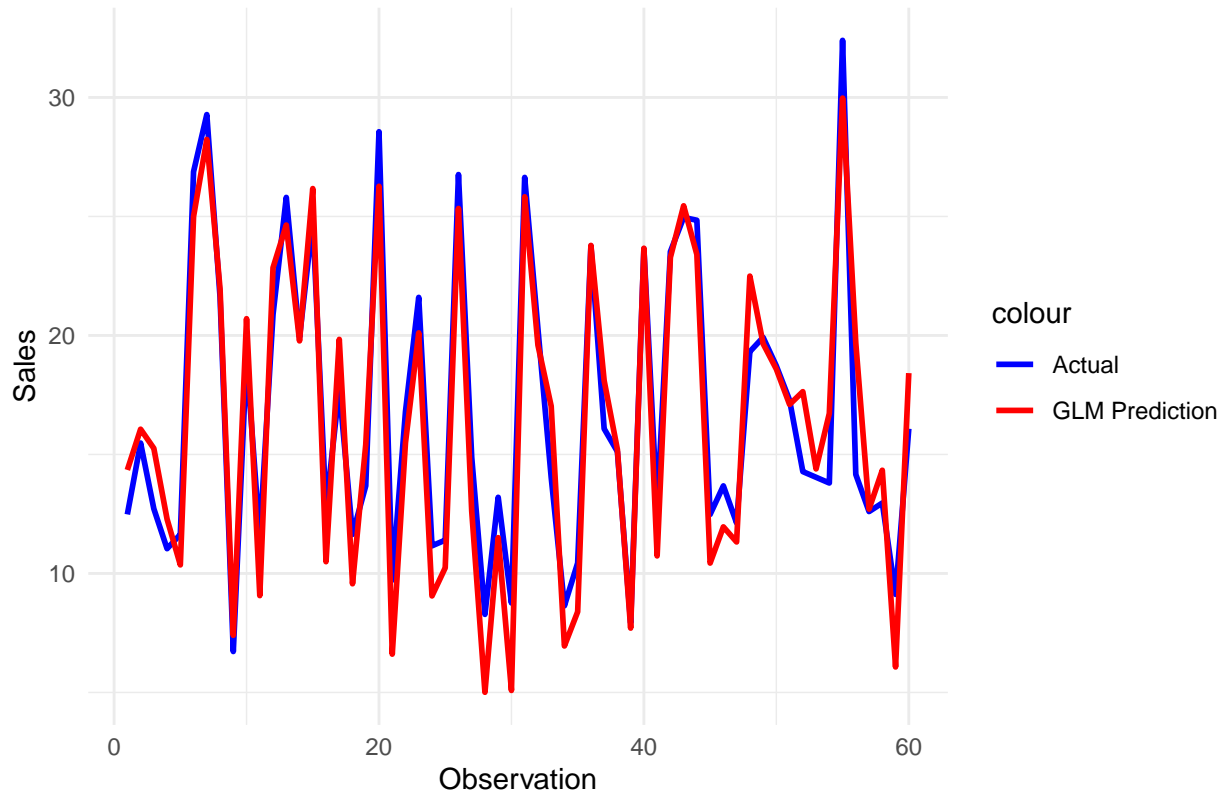
Comparison of Actual Sales and GLM Model Prediction



```
# Visualisasi Sales Aktual dengan Hasil Prediksi
ggplot(data = result_data_glm, aes(x = 1:length(sales))) +
  geom_line(aes(y = sales, color = "Actual"), size = 1) +
  geom_line(aes(y = Predictions, color = "GLM Prediction"), size = 1) +
  labs(x = "Observation", y = "Sales") +
  scale_color_manual(values = c("Actual" = "blue", "GLM Prediction" = "red")) +
  ggtitle("Comparison of Actual Sales and GLM Model Prediction") +
  theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

Comparison of Actual Sales and GLM Model Prediction



Konsep Evaluasi Model dengan Metrik Perfomansi Model: MAE, MSE, RMSE, dan MAPE:

1. Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Rata-rata dari nilai absolut dari selisih antara nilai sebenarnya y_i dan nilai prediksi \hat{y}_i . MAE memberikan gambaran tentang seberapa besar kesalahan prediksi rata-rata dalam skala asli.

2. Mean Squared Error (MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Rata-rata dari kuadrat selisih antara nilai sebenarnya y_i dan nilai prediksi \hat{y}_i . MSE memberikan lebih banyak bobot pada kesalahan besar.

3. Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left(\frac{|y_i - \hat{y}_i|}{|y_i|} \right) \times 100\%$$

Rata-rata persentase dari nilai absolut dari selisih antara nilai sebenarnya y_i dan nilai prediksi \hat{y}_i , dinormalkan dengan nilai sebenarnya. MAPE memberikan indikasi kesalahan relatif dalam persentase.

4. Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Akar kuadrat dari MSE. RMSE memberikan ukuran kesalahan yang serupa dengan MAE tetapi memberikan bobot lebih besar pada kesalahan yang lebih besar.

Pilihan metrik tergantung pada konteks dan preferensi spesifik. MAE, MSE, dan RMSE mengukur kesalahan dalam skala asli, sementara MAPE memberikan pengukuran kesalahan relatif dalam persentase. Dalam beberapa kasus, Anda mungkin perlu mempertimbangkan karakteristik spesifik dari masalah regresi yang sedang dihadapi untuk memilih metrik yang sesuai. Pemilihan metrik performansi tergantung pada tipe masalah dan tujuan spesifik dari model prediksi yang sedang dievaluasi.

Evaluasi Model

```
# Evaluasi GLM Model
library(Metrics)
```

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

```
# Hitung MAE (Mean Absolute Error)
mae_value_glm_1 <- mae(testingset$sales, predicted_glm)

# Hitung MSE (Mean Squared Error)
mse_value_glm_1 <- mse(testingset$sales, predicted_glm)

# Hitung RMSE (Root Mean Square Error)
rmse_value_glm_1 <- rmse(testingset$sales, predicted_glm)

# Hitung MAPE (Mean Absolute Percent Error)
mape_value_glm_1 <- mape(testingset$sales, predicted_glm)

# Tampilkan hasil evaluasi
cat(paste("MAE: ", mae_value_glm_1, "\n"))
```

```
## MAE: 1.60930616010998
```

```
cat(paste("MSE: ", mse_value_glm_1, "\n"))
```

```
## MSE: 3.8035119457981
```

```
cat(paste("RMSE: ", rmse_value_glm_1, "\n"))
```

```
## RMSE: 1.95025945602069
```

```
cat(paste("MAPE: ", mape_value_glm_1, "%\n"))
```

```
## MAPE:  0.117800954558218 %
```

Fit Model 2

```
# model GLM untuk memprediksi variabel sales pada data training
model_glm_2 <- glm(sales ~ youtube + facebook + newspaper, data = trainingset,
                  family = gaussian(link = "identity"), # distribusi probabilitas dari variabel respons
                  control = list(epsilon = 1e-8, # nilai batasan toleransi kesalahan (0.00000001)
                                maxit = 500) # jumlah iterasi 500 kali
                  )
```

```
# melihat hasil model prediksi
summary(model_glm_2)
```

```
##
## Call:
## glm(formula = sales ~ youtube + facebook + newspaper, family = gaussian(link = "identity"),
##      data = trainingset, control = list(epsilon = 1e-08, maxit = 500))
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.993025   0.472703   6.332 3.28e-09 ***
## youtube      0.048507   0.001722  28.170 < 2e-16 ***
## facebook     0.183034   0.011135  16.437 < 2e-16 ***
## newspaper    0.002454   0.007287   0.337  0.737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 4.359657)
##
## Null deviance: 5507.04  on 139  degrees of freedom
## Residual deviance:  592.91  on 136  degrees of freedom
## AIC: 609.38
##
## Number of Fisher Scoring iterations: 2
```

```
# Melakukan prediksi sales dengan model GLM pada data testing
predicted_glm_2 <- predict(model_glm_2, newdata = testingset, type = "response")

# Menampilkan hasil prediksi
head(predicted_glm_2)
```

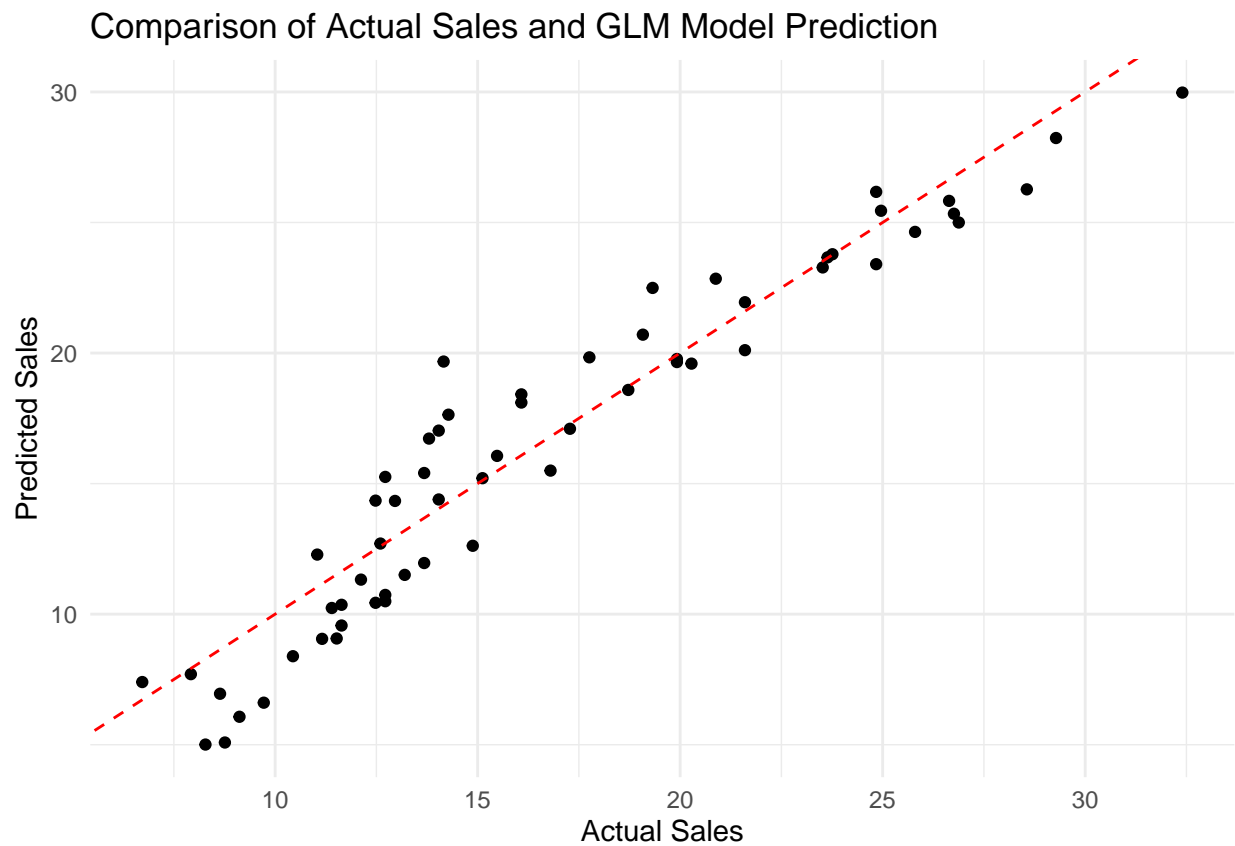
```
##           2           5           10           13           14           16
## 14.34799 16.06113 15.25647 12.28185 10.35878 24.99953
```

```
# Visualisasi Hasil
result_data_glm_2 <- data.frame(sales = testingset$sales, Predictions = predicted_glm_2)

# Visualisasi Perbandingan sales Aktual dengan Hasil Prediksi
```

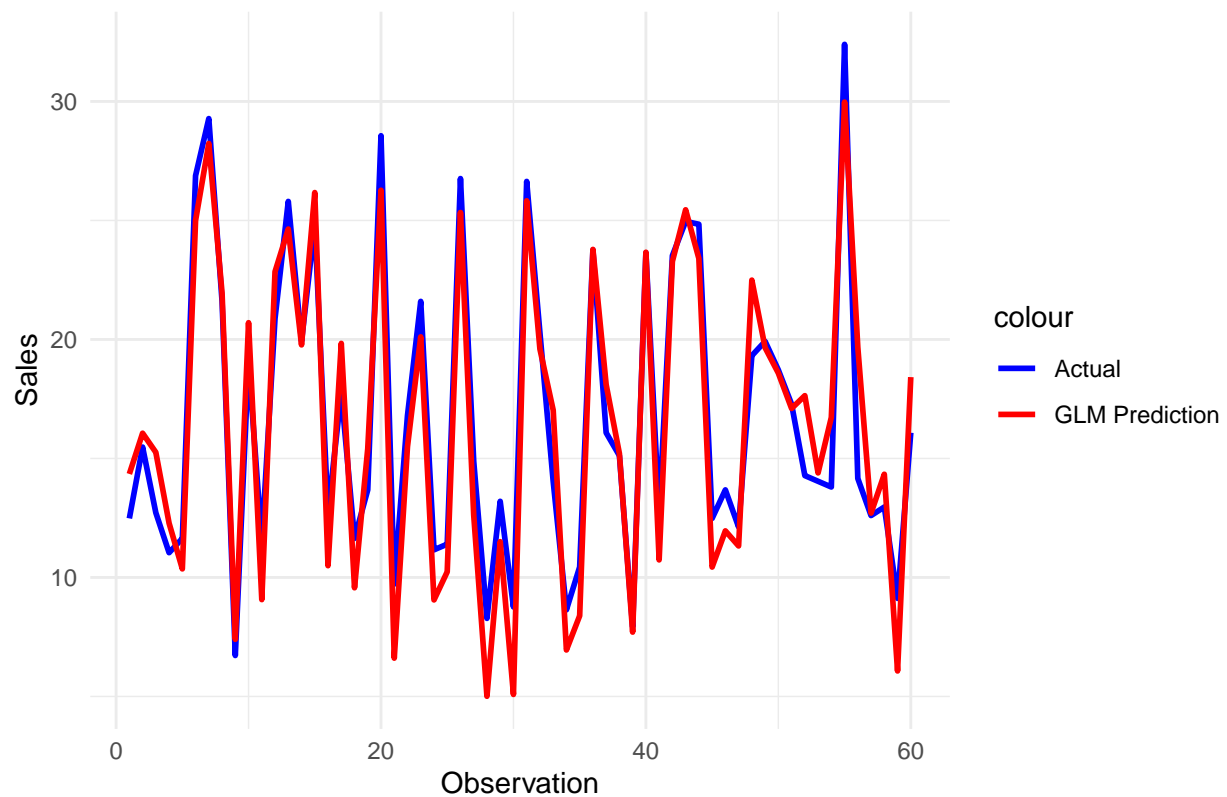


```
ggplot(data = result_data_glm_2, aes(x = sales, y = Predictions)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma) +
  labs(x = "Actual Sales", y = "Predicted Sales") +
  ggtitle("Comparison of Actual Sales and GLM Model Prediction") +
  theme_minimal()
```



```
# Visualisasi sales Aktual dengan Hasil Prediksi
ggplot(data = result_data_glm_2, aes(x = 1:length(sales))) +
  geom_line(aes(y = sales, color = "Actual"), size = 1) +
  geom_line(aes(y = Predictions, color = "GLM Prediction"), size = 1) +
  labs(x = "Observation", y = "Sales") +
  scale_color_manual(values = c("Actual" = "blue", "GLM Prediction" = "red")) +
  ggtitle("Comparison of Actual Sales and GLM Model Prediction") +
  theme_minimal()
```

Comparison of Actual Sales and GLM Model Prediction



```
# Evaluasi GLM Model

# Hitung MAE (Mean Absolute Error)
mae_value_glm_2 <- mae(testingset$sales, predicted_glm_2)

# Hitung MSE (Mean Squared Error)
mse_value_glm_2 <- mse(testingset$sales, predicted_glm_2)

# Hitung RMSE (Root Mean Square Error)
rmse_value_glm_2 <- rmse(testingset$sales, predicted_glm_2)

# Hitung MAPE (Mean Absolute Percent Error)
mape_value_glm_2 <- mape(testingset$sales, predicted_glm_2)

# Tampilkan hasil evaluasi
cat(paste("MAE: ", mae_value_glm_2, "\n"))
```

```
## MAE: 1.60930616010998
```

```
cat(paste("MSE: ", mse_value_glm_2, "\n"))
```

```
## MSE: 3.8035119457981
```

```
cat(paste("RMSE: ", rmse_value_glm_2, "\n"))
```

```
## RMSE:  1.95025945602069
```

```
cat(paste("MAPE: ", mape_value_glm_2, "%\n"))
```

```
## MAPE:  0.117800954558218 %
```