BAB IX
KORELASI DAN REGRESI LINEAR SEDERHANA

Capaian Pembelajaran

- Mahasiswa mengerti dan memahami hubungan linear antara kedua peubah dalam kaitannya dengan regresi linear sederhana adalah peubah bebas dan peubah tak bebas
- Mahasiswa mampu untuk menggambarkan tingkat keeratan hubungan linear antara dua peubah atau lebih
- Mahasiswa mampu membentuk model regresi linear sederhana dan menginterpretasikan model regresi yang dihasilkan
- Mahasiswa dapat menentukan kesesuaian model regresi yang dihasilkan.

9.1 Korelasi

Korelasi dan regresi adalah dua konsep penting dalam statistik yang sering digunakan untuk memahami hubungan antara dua atau lebih variabel. Korelasi mengukur hubungan linier antara dua variabel. Nilai korelasi berkisar antara -1 dan 1, di mana -1 menunjukkan hubungan negatif yang kuat, 1 menunjukkan hubungan positif yang kuat, dan 0 menunjukkan bahwa tidak ada hubungan antar dua variabel. Sedikit berbeda dengan korelasi, regresi juga digunakan untuk mengukur hubungan antar variabel, namun hubungan yang diukur adalah hubungan sebab akibat. Regresi memungkinkan kita untuk memprediksi nilai satu variabel dari nilai variabel lain. Ini biasanya dilakukan dengan membangun model matematis yang menggambarkan hubungan antara variabel tersebut. Dengan memahami konsep korelasi dan regresi dan bagaimana mereka berkaitan satu sama lain, metode regresi dapat digunakan untuk memahami dan memprediksi hubungan antara variabel dalam berbagai bidang, seperti bisnis, ekonomi, ilmu sosial, dan lainnya.

Rumus korelasi antara dua variabel X dan Y adalah:

$$r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][\sum Y^2 - (\sum Y)^2]}}$$

di mana

n : Jumlah observasi,

X dan Y : Nilai dari masing-masing variabel,

ΣX dan ΣY : Jumlah dari masing-masing variabel,

ΣXY : Jumlah produk dari X dan Y,

ΣX^2 dan ΣY^2 : Jumlah kuadrat dari masing-masing variabel.

9.2 Persamaan Regresi Linier Sederhana

Dalam regresi linier sederhana, model adalah garis yang menghubungkan titik-titik data. Garis sederhana hanya cocok untuk memodelkan hubungan linier antara dua variabel. Dalam kasus di mana hubungan antara variabel tidak linier, regresi non-linier atau regresi polinomial mungkin lebih cocok. Dalam regresi berganda, kita dapat memodelkan hubungan antara lebih dari dua variabel. Ini bisa digunakan untuk memprediksi variabel tergantung dari lebih dari satu variabel independen. Sebelum melakukan regresi, penting untuk melakukan analisis korelasi untuk memastikan bahwa ada hubungan antara variabel yang ingin kita modelkan. Jika tidak ada hubungan, maka regresi akan memberikan hasil yang tidak berguna. Dalam analisis regresi, ada beberapa asumsi yang harus dipenuhi agar hasil analisis dapat diinterpretasikan dengan benar. Salah satu asumsi yang paling penting adalah asumsi linearitas, yaitu bahwa hubungan antara variabel harus linier.Asumsi lain meliputi homoskedastisitas (varian yang sama untuk setiap nilai dari variabel independen), independensi observasi/non-multikolinieritas (tidak ada hubungan antar observasi), dan normalitas (distribusi residual harus berdistribusi normal). Setelah asumsi dipenuhi, model regresi dapat dibuat dan diuji untuk melihat apakah memenuhi syarat.

Persamaan model regresi linier sederhana adalah:

$$Y = b_0 + b_1 X$$

di mana:

b_0 : Intersep,

b_1 : Koefisien regresi,

X : Variabel independen.

Koefisien regresi, $b1$, dapat ditemukan dengan rumus:

$$b_i = r \times \frac{s_Y}{s_X}$$

di mana:

r : Korelasi antara X dan Y,

S_Y : Standar deviasi dari Y, dan

S_X : Standar deviasi dari X.

Intersep, b_0, dapat ditemukan dengan rumus:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

di mana:

$\bar{Y}$: rata-rata dari Y,

$\bar{X}$: rata-rata dari X.

Dengan menggunakan rumus ini, kita dapat membuat model regresi dan memprediksi nilai dari variabel dependen berdasarkan nilai variabel independen.

Uji Serempak dan Uji Individual

Uji serempak adalah uji hipotesis yang menguji apakah seluruh variabel independen secara bersama-sama memiliki pengaruh signifikan terhadap variabel dependen.

$H0$: $\beta_1 = 0$ (Variabel bebas tidak mempengaruhi variabel tak bebas),

$H1$: $\beta_1 \neq 0$ (Variabel bebas memiliki pengaruh terhadap variabel tak bebas).

Statistik *F* digunakan untuk uji serempak, statistik *F* dihitung dengan membagi varians dari *mean square error* (MSE) dengan varians *mean square* dari model (MSM).

Uji parsial adalah uji hipotesis yang menguji apakah setiap variabel independen secara terpisah memiliki pengaruh signifikan terhadap variabel dependen.

$H0$: $\beta_1 = 0$ (tidak ada hubungan signifikan antara variabel independen dan dependen)

$H1$: $\beta_1 \neq 0$ (ada hubungan signifikan antara variabel independen dan dependen)

Statistik *t* digunakan untuk uji parsial. Statistik *t* dihitung dengan membagi nilai estimasi dari koefisien regresi oleh standar deviasi dari estimasi. Kedua jenis uji hipotesis ini memiliki tujuan yang berbeda, tetapi hasil dari kedua uji ini seringkali digabungkan untuk memperoleh gambaran yang lebih komprehensif tentang pengaruh variabel independen terhadap variabel dependen. Kedua staistik uji (uji *F* dan uji *t*) dibandingkan dengan nilai-nilai dari tabel *F* atau *t* pada tingkat signifikansi tertentu untuk memutuskan apakah hipotesis nol dapat diterima atau ditolak. Jika nilai statistik lebih besar dari nilai tabel, maka hipotesis nol ditolak dan hipotesis alternatif diterima, yang berarti ada pengaruh signifikan antara variabel independen dan dependen.

9.3 Aplikasi Menggunakan R

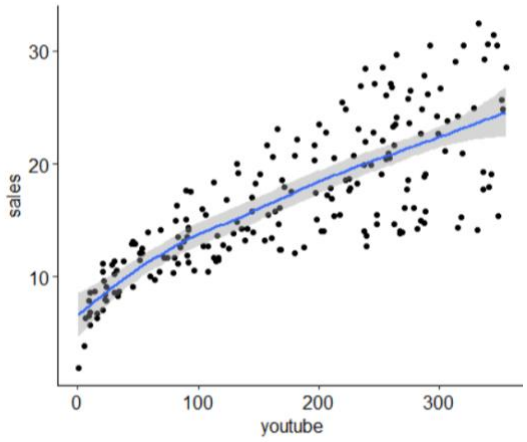
Terdapat banyak dataset yang disediakan oleh *R* untuk diuji hubungannya menggunakan metode korelasi dan dimodelkan menggunakan regresi. Salah satunya adalah *marketing dataset* dalam *package* `datarium` yang memuat informasi mengenai dampak dari tiga media periklanan (*youtube*, *facebook*, dan surat kabar) terhadap penjualan. Data terdiri dari data anggaran iklan (variabel independen) dan data penjualan (variabel dependen) dalam ribuan dolar. Berikut adalah `summary` dari data.

```
> # Load the package
> data("marketing", package = "datarium")
> head(marketing, 4)
```

```
  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
```

Sebelum melakukan analisis korelasi dan regresi, sebaiknya dilihat dulu pola hubungan antar variabel melalui *scatter plot*. Berikut adalah pola hubungan antara variabel penjualan dan youtube.

```
#scatterplot
ggplot(marketing,aes(x=youtube,y=sales))+geom_point()+ stat_smooth())
```



Grafik di atas menunjukkan bahwa terdapat hubungan yang linier antara penjualan dan youtube. Artinya, apabila iklan di youtube meningkat, maka penjualan juga ikut meningkat. Hasil ini cukup baik karena salah satu asumsi dalam regresi adalah linieritas. Nilai korelasi antara kedua variabel tersebut dapat dihitung menggunakan perintah berikut:

```
> cor(marketing$sales, marketing$youtube)
[1] 0.7822244
```

Nilai koefisien korelasi antara variabel penjualan dan youtube adalah 0,78, artinya terdapat hubungan yang arahnya positif dan cukup kuat antara variabel penjualan dan youtube.

Model regresi linier dapat dibangun menggunakan perintah `lm()` sebagai berikut:

```
> #regresi
> model <- lm(sales ~ youtube, data = marketing)
> model

Call:
lm(formula = sales ~ youtube, data = marketing)

Coefficients:
(Intercept)      youtube
```

```
8.43911       0.04754
```

Model regresi yang dihasilkan adalah

$$Y = 8{,}43911 + 0{,}04754X$$

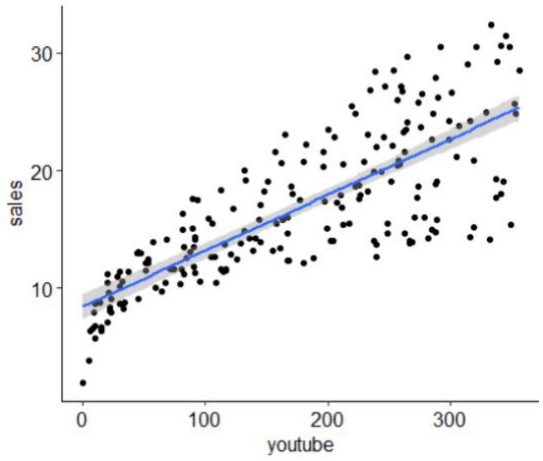
Interpretasi:

- Nilai intersep 8,44 menunjukkan bahwa penjualan akan bernilai sebesar 8,44 ribu dolar jika tidak ada iklan di youtube. Artinya, jika iklan di youtube sama dengan nol, maka penjualan akan senilai 8,44 * 1000 = 84440 dolar.
- Nilai koefisien regresi untuk variabel youtube adalah sebesar 0,048. Artinya, apabila biaya iklan di youtube sama dengan 1000 dolar, maka akan meningkatkan penjualan sebesar 48 unit (0,048 * 1000). Sehingga penjualan akan menjadi sales = 8.44 + 0.048*1000 = 56.44 unit atau 56440 dolar.

Garis regresi dapat ditambahkan ke scatter plot dengan perintah stat_smooth() [ggplot2].

```
#garis regresi
ggplot(marketing, aes(youtube, sales)) + geom_point() +
stat_smooth(method = lm)
```

Diperoleh hasil garis regresi disekitar titik data sebagai berikut:



Summary model regresi untuk menampilkan nilai koefisen determinasi (R^2) dan hasil pengujian hipotesis secara serentak dan parsial. Output dari `summary(model)` terdiri dari enam komponen yaitu:

- **Call**, menunjukkan fungsi yang dipanggil untuk menghitung model regresi.
- **Residual**, menunjukkan distribusi dari residual (selisih titik data dan garis regresi).
- **Coefficient**, menunjukkan nilai koefisien beta dan signifikansinya. Variabel independen yang signifikan mempengaruhi variabel dependen dapat dilihat dari nilai *p-value* yang diberi tanda bintang (***).

- **Residual standard error** (RSE), **R-Squared** (R^2), dan **statistik uji F** yang digunakan untuk mengukur kebaikan model.

Hipotesis yang diuji adalah:

$H0$: $\beta 1 = 0$ (Variabel youtube tidak mempengaruhi variabel penjualan),

$H1$: $\beta 1 \neq 0$ (Variabel youtube mempengaruhi variabel penjualan).

Berikut adalah *output* `summary(model)` dari data marketing:

```
> summary(model)
Call:
lm(formula = sales ~ youtube, data = marketing)
Residuals:
     Min       1Q   Median       3Q      Max
-10.0632  -2.3454  -0.2295   2.4805   8.6548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.439112   0.549412   15.36   <2e-16 ***
youtube     0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.91 on 198 degrees of freedom
Multiple R-squared:  0.6119,  Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Berdasarkan output di atas, diketahui bahwa baik pada pengujian hipotesis secara serentak maupun parsial, variabel youtube signifikan mempengaruhi penjualan karena *p-value* < α (5%). Nilai R^2 sebesar 0,6119 menunjukkan bahwa variabel youtube dapat menjelaskan variabel penjualan sebesar 61,19% sedangkan sisanya dijelaskan oleh variabel lain yang tidak masuk kedalam model.

9.4 Latihan

1. Diberikan sintaks R sebagai berikut

```
# Memuat data
data(mtcars)
# Membuat model regresi linear sederhana
```

```
fit <- lm(mpg ~ wt, data = mtcars)
# Melihat summary dari model
summary(fit)
# Plot data dan garis regresi
library(ggplot2)
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

Dalam contoh di atas, kita memuat data mtcars dari paket built-in R. Kemudian, kita membuat model regresi linear sederhana dengan menggunakan fungsi lm() dan memasukkan variabel dependen mpg dan variabel independen wt. Kita juga menggunakan fungsi summary() untuk melihat ringkasan dari model dan memplot data beserta garis regresi menggunakan paket ggplot2.

2. Dalam suatu produk, tingkat air dalam campuran basah (X) diprediksi memiliki pengaruh terhadap kepadatan produk akhir (Y). Dalam suatu eksperiment, tingkat air dalam campuran dikontrol dan kepadatan produk akhir diukur. Hasil data yang terkumpul adalah sebagai berikut:

Y	X
3	4.7
3	5.0
4	5.2
5	5.9
10	5.3
2	5.6
9	5.0
3	4.7
7	5.0

a. Hitung koefisien korelasi. Adakah hubungan antara kadar air campuran basah dan kepadatan produk?
b. Lakukan pengujian terhadap koefisien korelasi
c. Tentukan persamaan garis regresinya