

# Modul IS240 Lab 11: Korelasi (Rev 2)

Tan Thing Heng (Prodi Sistem Informatika UMN)

```
library(readxl)
library(MVN)
library(tidyverse)

food <- read_excel("data/labdata.xlsx", sheet = "ConsumerFood")
advertising <- read_excel("data/ProbStatDataSet.xlsx", sheet = "Black12_8Advertising")
lamb <- read_excel("data/ProbStatDataSet.xlsx", sheet = "BlackTable17_6Lamb")
index <- read_excel("data/ProbStatDataSet.xlsx", sheet = "BlackTable12_1Index")

food$Region <- as.factor(food$Region)
food$Location <- as.factor(food$Location)
str(food)

## Classes 'tbl_df', 'tbl' and 'data.frame': 200 obs. of 5 variables:
## $ AnnualFoodSpending : num 8909 5684 10706 14112 13855 ...
## $ AnnualHouseholdIncome : num 56697 35945 52687 74041 63182 ...
## $ NonMortgageHouseholdDebt: num 23180 7052 16149 21839 18866 ...
## $ Region : Factor w/ 4 levels "NE","NW","S",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ Location : Factor w/ 2 levels "Metro","NonMetro": 1 1 1 1 1 1 1 1 1 1 ...
```

## 1 Pendahuluan

### 1.1 Tujuan Pembelajaran

- Mahasiswa mampu menganalisa hubungan antar variabel pada data.

### 1.2 Topik Praktikum

- Korelasi Pearson
- Korelasi Spearman

### 1.3 Asosiasi, Korelasi dan Kausasi

- Asosiasi adalah hubungan antara dua buah variabel (independen atau dependen).
- Korelasi adalah kekuatan hubungan linier antara 2 buah variabel numerik.
- $cor(X, Y) = cor(Y, X)$
- Kausasi (**causation**) adalah hubungan sebab akibat.
- Korelasi **tidak sama** dengan kausalitas. Koefisien korelasi yang tinggi hanya menunjukkan kuatnya suatu hubungan, **bukan** arah kausalitas.

- **Spurious correlation** terjadi bila terdapat korelasi yang kuat antara 2 variabel numerik yang tidak ada hubungannya. Contoh: korelasi antara penjualan es krim dengan skor kuis Statistika.

## 1.4 Tipe Asosiasi

Uji	Tipe Variabel	Syarat Penggunaan
Pearson	Numerik & Numerik	data dari sampel independen dan acak, distribusi data bivariat Gaussian
Spearman	Numerik & Numerik	data dari sampel independen dan acak, distribusi data <b>tidak</b> bivariat Gaussian
Spearman	Kategorikal & Kategorikal	data dari sampel independen dan acak, skala pengukuran ordinal
Kendall	Kategorikal & Kategorikal	data dari sampel independen dan acak, s.p. ordinal, bila tdp ( <i>tied rank</i> )
Chi-Kuadrat	Kategorikal & Kategorikal	data dari sampel independen dan acak, s.p. nominal

Tipe asosiasi yang dibahas pada pertemuan ini adalah korelasi Pearson dan korelasi Spearman.

## 1.5 Visualisasi Korelasi dengan Diagram Acak

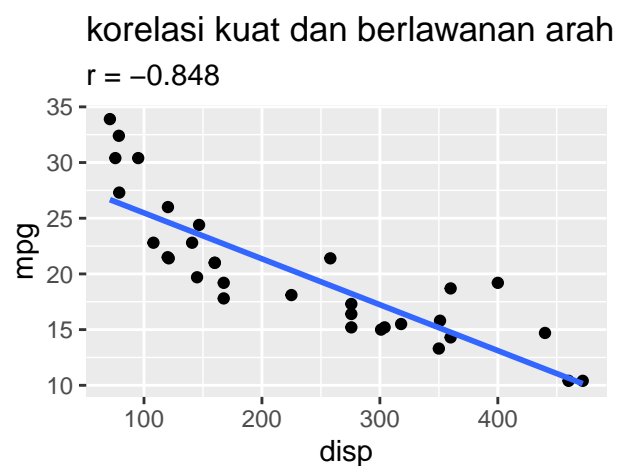
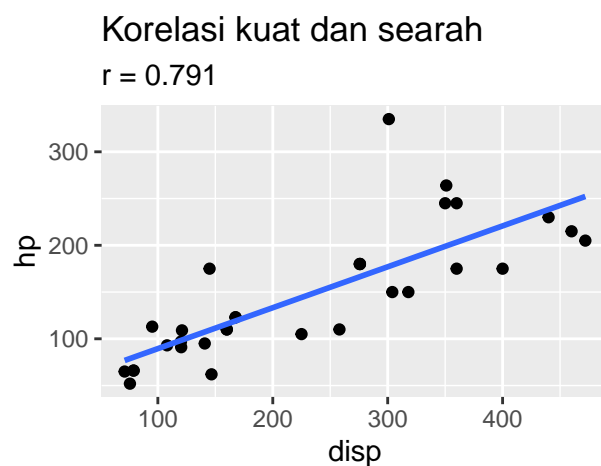
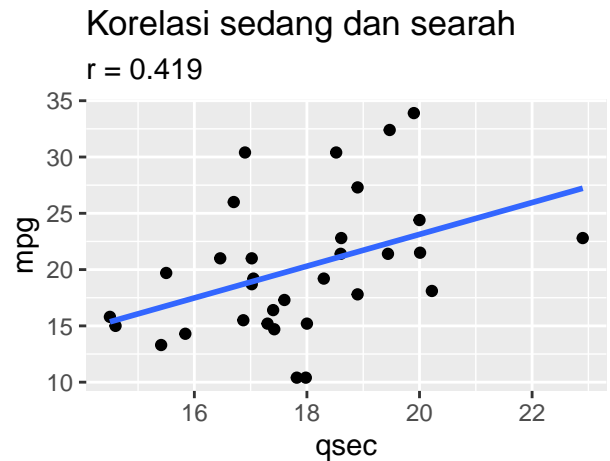
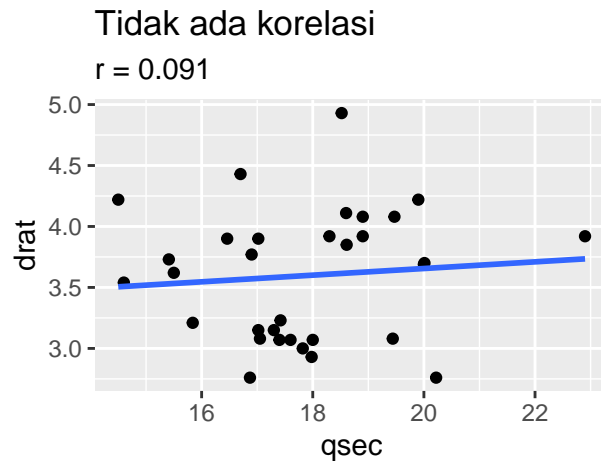
```
p1 <- mtcars %>% ggplot(aes(qsec, drat))+ geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Tidak ada korelasi", subtitle = paste("r =",round(cor(mtcars$qsec, mtcars$drat),3)))

p2 <- mtcars %>% ggplot(aes(qsec, mpg))+ geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Korelasi sedang dan searah", subtitle = paste("r =",round(cor(mtcars$qsec, mtcars$mpg),3)))

p3 <- mtcars %>% ggplot(aes(dis, hp))+ geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Korelasi kuat dan searah", subtitle = paste("r =",round(cor(mtcars$dis, mtcars$hp),3)))

p4 <- mtcars %>% ggplot(aes(dis, mpg))+ geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("korelasi kuat dan berlawanan arah", subtitle = paste("r =",round(cor(mtcars$dis, mtcars$mpg),3)))

gridExtra::grid.arrange(p1, p2, p3, p4, nrow = 2)
```



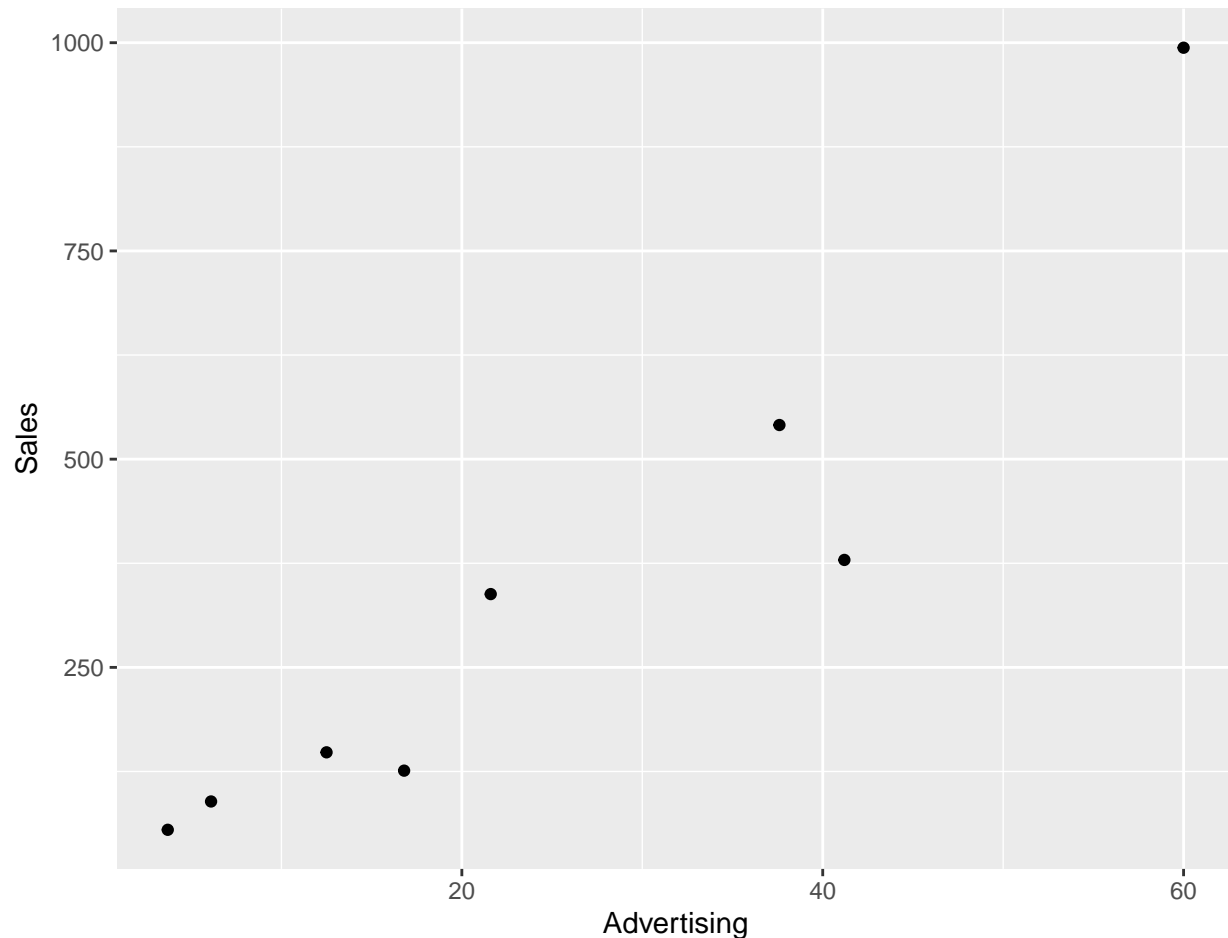
## 2 Contoh Korelasi Pearson

Gunakan uji hipotesa yang sesuai untuk menguji korelasi antara Advertising dan Sales. Gunakan tingkat signifikansi 5%.

advertising

```
## # A tibble: 8 x 2
##   Advertising Sales
##   <dbl> <dbl>
## 1    12.5   148
## 2     3.7    55
## 3    21.6  338
## 4     60   994
## 5    37.6  541
## 6     6.1   89
## 7    16.8  126
## 8    41.2  379

# plot(advertising$Advertising, advertising$Sales)
ggplot(advertising, aes(Advertising, Sales)) + geom_point()
```



Dari scatter plot terlihat hubungan yang kuat antara Sales dan Advertising.

```
library(MVN) #untuk uji normalitas multivariat
result <- mvn(advertising, mvnTest = "royston")
names(result)
```

```
## [1] "multivariateNormality" "univariateNormality" "Descriptives"
```

```
result$multivariateNormality
```

```
##      Test      H    p value MVN
## 1 Royston 2.318991 0.1838085 YES
```

```
result$univariateNormality
```

```
##      Test      Variable Statistic    p value Normality
## 1 Shapiro-Wilk Advertising    0.9225    0.4502      YES
## 2 Shapiro-Wilk    Sales        0.8431    0.0811      YES
```

Hasil uji normalitas multivariate dengan uji Royston menunjukkan data Advertising bivariat normal. Uji normalitas univariat juga menunjukkan normalitas untuk kedua variabel. Kita dapat menggunakan uji korelasi Pearson.

```
ans <- cor.test(advertising$Advertising, advertising$Sales, method = "pearson", conf.level = 0.95)
ans
```

```
##
```

```
## Pearson's product-moment correlation
##
## data: advertising$Advertising and advertising$Sales
## t = 7.2705, df = 6, p-value = 0.0003445
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## 0.7314386 0.9907323
## sample estimates:
## cor
## 0.9476626
```

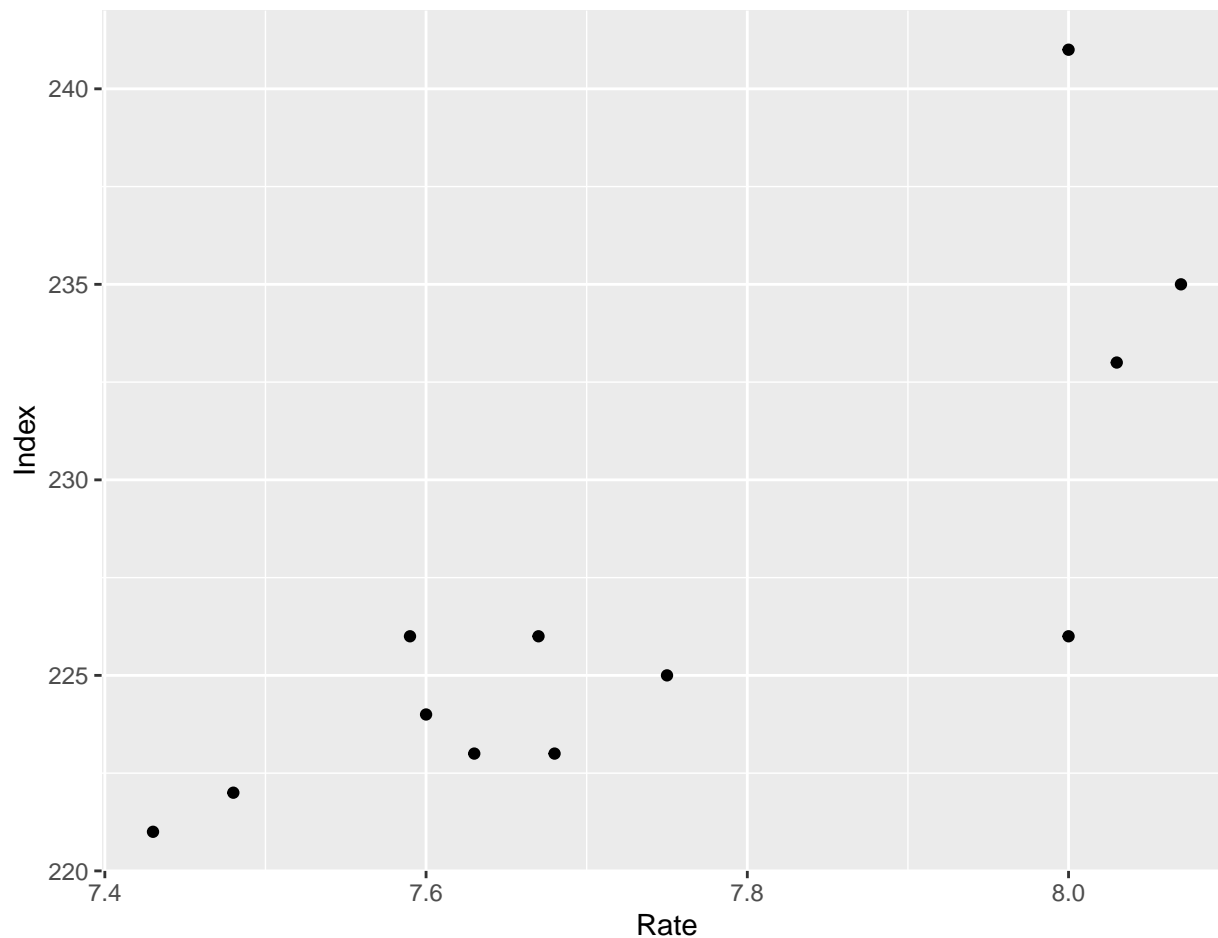
Hasil uji korelasi linier Pearson signifikan ( $p\text{-value} = 3.4448648 \times 10^{-4} < 0,05$ ). Koefisien korelasi  $r = 0.9476626$  menunjukkan hubungan linier yang sangat kuat antara Advertising dan Sales. Makin tinggi biaya untuk Advertising, makin tinggi Sales.

### 3 Contoh Korelasi Spearman

```
index
```

```
## # A tibble: 12 x 2
##   Rate Index
##   <dbl> <dbl>
## 1  7.43  221
## 2  7.48  222
## 3   8    226
## 4  7.75  225
## 5  7.6    224
## 6  7.63  223
## 7  7.68  223
## 8  7.67  226
## 9  7.59  226
## 10 8.07  235
## 11 8.03  233
## 12 8     241
```

```
# plot(index$Rate, index$Index)
ggplot(index, aes(Rate, Index)) + geom_point()
```



```
library(MVN) #untuk uji normalitas multivariat
result <- mvn(index, mvnTest = "royston")
names(result)
```

```
## [1] "multivariateNormality" "univariateNormality" "Descriptives"
```

```
result$multivariateNormality
```

```
##      Test      H      p value MVN
## 1 Royston 7.178837 0.02335825 NO
```

```
result$univariateNormality
```

```
##      Test Variable Statistic p value Normality
## 1 Shapiro-Wilk Rate      0.8948 0.1358 YES
## 2 Shapiro-Wilk Index     0.8279 0.0198 NO
```

Hasil uji normalitas multivariate dengan uji Royston menunjukkan data ini **tidak** bivariat normal. Uji normalitas univariat untuk Rate normal namun uji normalitas univariate tidak menunjukkan normalitas untuk Index. Kita dapat menggunakan uji korelasi Spearman.

```
ans <- cor.test(index$Rate, index$Index, method = "spearman", conf.level = 0.95)
ans
```

```
##
## Spearman's rank correlation rho
```

```
##
## data: index$Rate and index$Index
## S = 59.62, p-value = 0.002159
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.7915392
```

Hasil uji korelasi linier Spearman signifikan ( $p\text{-value} = 0.0021588 < 0,05$ ). Koefisien korelasi  $r = 0.7915392$  menunjukkan hubungan linier yang sangat kuat antara Rate dan Index. Makin tinggi Rate, makin tinggi Future Index.

Bila kita hanya mengetahui ranking Index dan Rate (tidak mengetahui nilai Rate dan nilai Index), kita juga menggunakan korelasi Spearman untuk menghitung kekuatan korelasinya.

```
index$RateRank <- rank(index$Rate, ties.method = "average")
index$IndexRank <- rank(index$Index, ties.method = "average")
index$d <- index$RateRank - index$IndexRank
index
```

```
## # A tibble: 12 x 5
##   Rate Index RateRank IndexRank    d
##   <dbl> <dbl>   <dbl>   <dbl> <dbl>
## 1  7.43   221     1       1     0
## 2  7.48   222     2       2     0
## 3    8    226    9.5     8    1.5
## 4  7.75   225     8       6     2
## 5  7.6    224     4       5    -1
## 6  7.63   223     5      3.5    1.5
## 7  7.68   223     7      3.5    3.5
## 8  7.67   226     6       8    -2
## 9  7.59   226     3       8    -5
## 10 8.07   235    12      11     1
## 11 8.03   233    11      10     1
## 12 8      241    9.5      12   -2.5
```

Selisih rank  $d$  kemudian dapat digunakan untuk menghitung koefisien korelasi Spearman seperti pada halaman 706.

```
n <- NROW(index)
d <- index$d

r <- 1 - 6 * sum(d^2)/(n * n^2 - 1)
r

## [1] 0.7950203
```