Benefits and Risks of Correlations in Decision Making Using R for Analysis

Chihiro Sato (400924769)

2025-12-10

Before we start, please visit my GitHub via QR code or link

Agenda

- What is correlation?
- Benefits of using correlations
- Risks of using correlations
- Benefits and risks of using R
- Example 1: Benefit
- Example 2: Risk
- Exercise

What is Correlation?

- A statistical measure that shows how strongly and in which way two or more variables are related.
- Example: “temperature and ice cream sales,” “disposable income and expenditure for luxury items,” or “customer services and customer complaints”

Correlation coefficients

- Ranging from -1 to +1
- **Pearson** measures the strength of a linear relationship between two continuous variables
- **Spearman** measures the strength of a monotonic relationship, which can be linear or curved. It uses ranks to calculate
- **Kendall** measures the strength of dependence between two variables, also for ordinal data.

Comparison between Pearson and Spearman

Feature	Pearson	Spearman
Relationship	Linear	Monotonic (linear or not)
Data type	Continuous	Continuous or ordinal
Outliers	Sensitive	Robust
Normality	Assumes approx normal	No strict assumption

Role of correlations in decision-making

- When making decisions, humans have limited cognitive capacity, incomplete information, and face uncertainty.
- In business or research, correlation helps us understand which factors might influence outcomes. A high correlation suggests a strong relationship, which can guide decisions.

why use R? - R is a programming language specialized for statistical analysis. It makes correlation analysis and visualization (graphs) easy. Using R will make your decision-making more convincing.

Benefits of using correlations

- **Identifying Relationships**: Correlation analysis helps identify how two variables are interconnected.
- **Prediction and Forecasting**: Based on the relationship identified through past correlations, you can make predictions and correlations about future outcomes.
- **Strategy Optimization**: Analyzing historical data allows you to understand which initiatives have been most effective and enables you to formulate strategies accordingly.
- **Resource Allocation**: By understanding past effective and efficient practices, you can focus on impactful variables.

Risks of using correlations

- **Correlation Is Not Causation**: Just because two variables are correlated, it does not mean one causes the other; Misinterpretation can lead to wrong decisions.
- **External Factors**: It is necessary to consider external influences (or seasonal trends) that might impact the observed relationship between variables.
- **False Correlations**: Random data which actually have no relationship may accidentally look meaningful and lead you to wrong conclusions.

Benefits and risks of using R

Benefits of using R

- Powerful statistical capabilities
- Extensive package ecosystem
- Strong data visualization capabilities

Risks of using R

- Correlation ≠ Causation
- Complexity for non-experts
- False sense of precision

Example 1: Benefit

dataset “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

Correlation 4x4 matrix (Pearson = linear relationship)

```r
M <- cor(marketing, use = "pairwise.complete.obs", method = "pearson")
round(M, 3)
```

```
           youtube facebook newspaper sales
youtube      1.000    0.055     0.057 0.782
facebook     0.055    1.000     0.354 0.576
newspaper    0.057    0.354     1.000 0.228
sales        0.782    0.576     0.228 1.000
```

Correlation between two specific variables

```r
cor(marketing$youtube, marketing$sales)
```

```
[1] 0.7822244
```

```r
cor.test(marketing$youtube, marketing$sales)
```

```
    Pearson's product-moment correlation

data:  marketing$youtube and marketing$sales
t = 17.668, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7218201 0.8308014
sample estimates:
      cor
0.7822244
```

Spearman alternative = monotonic relationship

```r
M_s <- cor(marketing, use = "pairwise.complete.obs", method = "spearman")
round(M_s, 3)
```

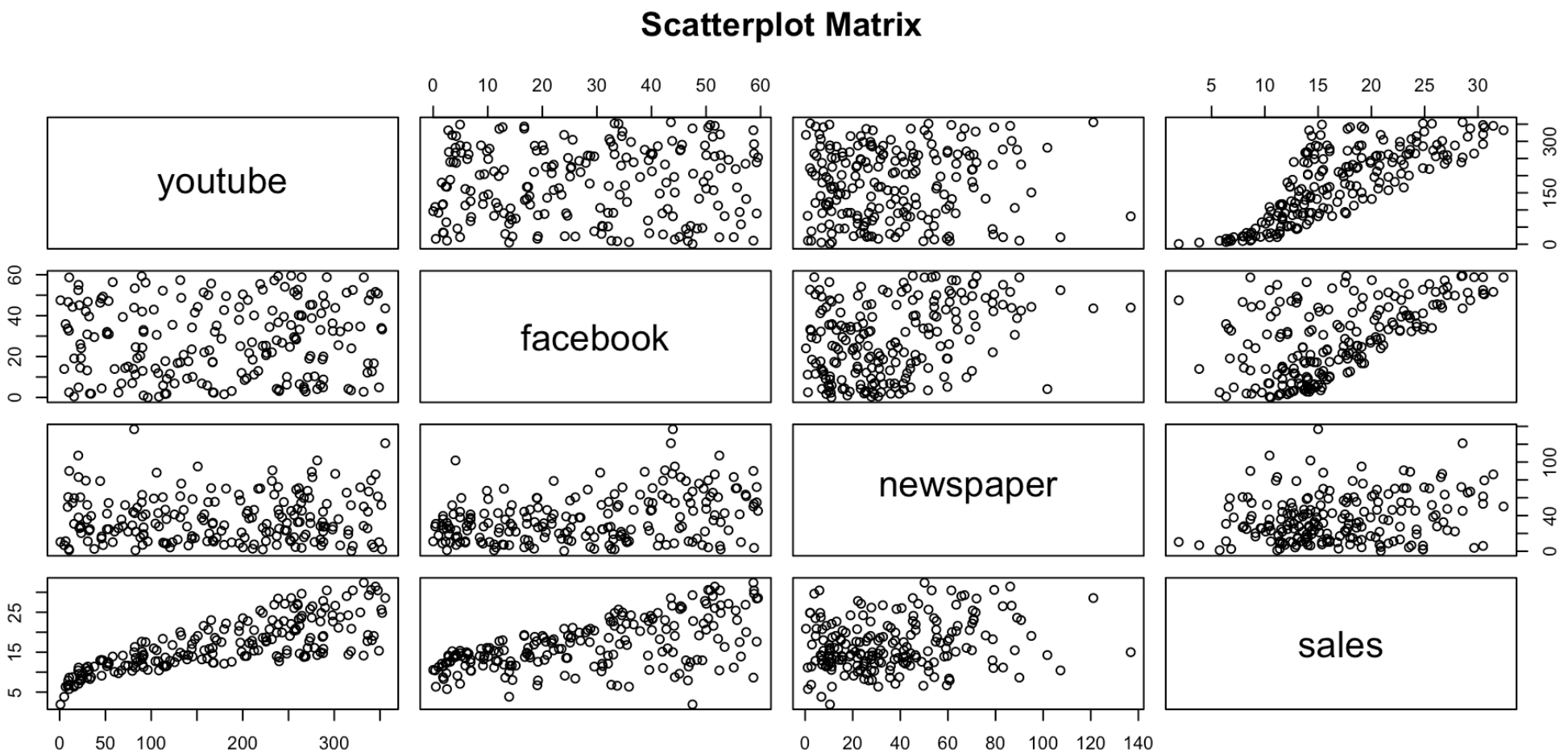
```
           youtube facebook newspaper sales
youtube      1.000    0.056     0.051 0.801
facebook     0.056    1.000     0.317 0.554
newspaper    0.051    0.317     1.000 0.195
sales        0.801    0.554     0.195 1.000
```

Visualization (heatmap)

```r
corrplot(M, method = "color", addCoef.col = "black", tl.col = "black")
```

Visualization (Scatterplot)

```
1 pairs(marketing, main = "Scatterplot Matrix")
```



Example 2: Risk

dataset “airquality”

```
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

```r
aq <- na.omit(airquality[, c("Ozone", "Solar.R", "Temp")])
```

Naive correlation (Ozone vs Solar.R)

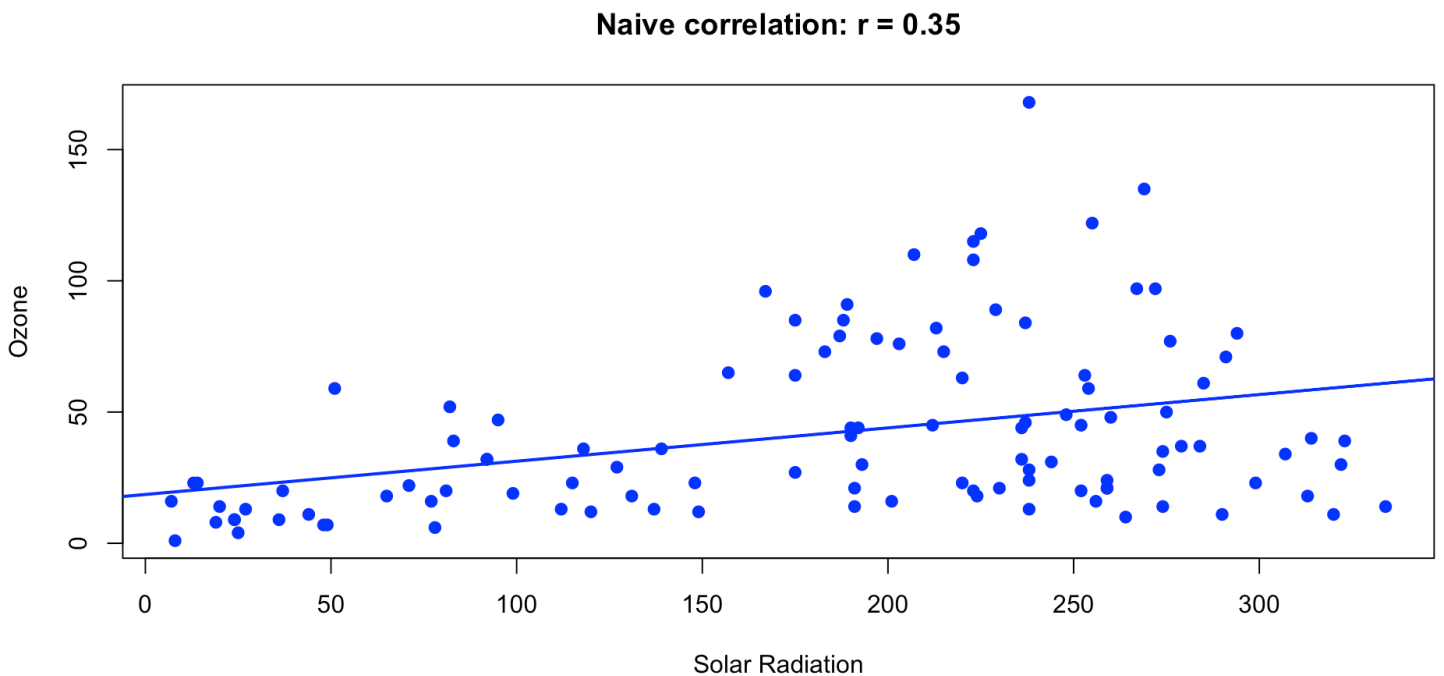
```r
cor(aq$Ozone, aq$Solar.R)
```

```
[1] 0.3483417
```

```r
r_naive <- cor(aq$Ozone, aq$Solar.R)
```

Visualization of naive correlation

```r
plot(aq$Solar.R, aq$Ozone,
     pch = 19, col = "blue",
     xlab = "Solar Radiation",
     ylab = "Ozone",
     main = "Naive correlation: r = 0.35")
abline(lm(Ozone ~ Solar.R, data = aq), col = "blue", lwd = 2)
```



Partial correlation using ppcor

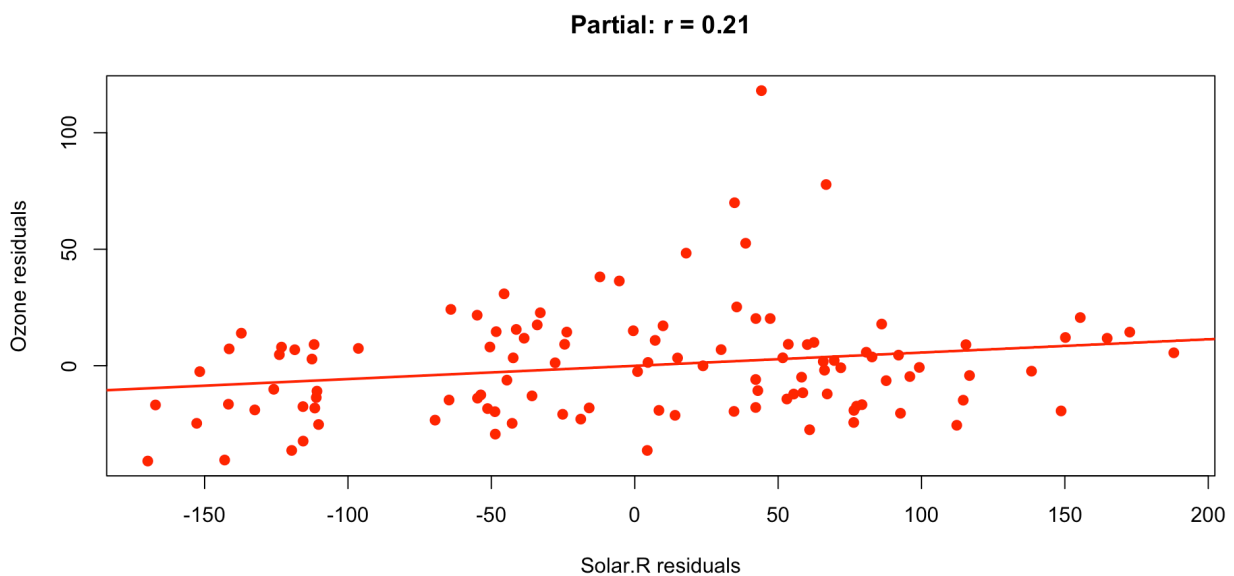
```r
library(ppcor)
pcor.test(x = aq$Ozone, y = aq$Solar.R, z = aq["Temp"])
```

```
   estimate    p.value statistic   n gp  Method
1 0.2089543 0.02847063  2.220534 111  1 pearson
```

```r
pc <- pcor.test(x = aq$Ozone, y = aq$Solar.R, z = aq["Temp"])
```

Visualization of partial correlation

```r
plot(residuals(lm(Solar.R ~ Temp, data = aq)),
     residuals(lm(Ozone ~ Temp, data = aq)),
     pch = 19, col = "red",
     xlab = "Solar.R residuals",
     ylab = "Ozone residuals",
     main = "Partial: r = 0.21")
abline(lm(residuals(lm(Ozone ~ Temp, data = aq)) ~ residuals(lm(Solar.R ~ T
            col = "red", lwd = 2)
```



Exercise

```r
data(anscombe)
head(anscombe)
```

```
  x1 x2 x3 x4   y1   y2    y3   y4
1 10 10 10  8 8.04 9.14  7.46 6.58
2  8  8  8  8 6.95 8.14  6.77 5.76
3 13 13 13  8 7.58 8.74 12.74 7.71
4  9  9  9  8 8.81 8.77  7.11 8.84
5 11 11 11  8 8.33 9.26  7.81 8.47
6 14 14 14  8 9.96 8.10  8.84 7.04
```

1. Calculate the correlation for each pair

2. Plot the data

Possible Solution

1. Calculate the correlation for each pair

```r
B <- cor(anscombe, use = "pairwise.complete.obs", method = "pearson")
round(B, 3)
```

```
      x1     x2     x3     x4     y1     y2     y3     y4
x1  1.000  1.000  1.000 -0.500  0.816  0.816  0.816 -0.314
x2  1.000  1.000  1.000 -0.500  0.816  0.816  0.816 -0.314
x3  1.000  1.000  1.000 -0.500  0.816  0.816  0.816 -0.314
x4 -0.500 -0.500 -0.500  1.000 -0.529 -0.718 -0.345  0.817
y1  0.816  0.816  0.816 -0.529  1.000  0.750  0.469 -0.489
y2  0.816  0.816  0.816 -0.718  0.750  1.000  0.588 -0.478
y3  0.816  0.816  0.816 -0.345  0.469  0.588  1.000 -0.155
y4 -0.314 -0.314 -0.314  0.817 -0.489 -0.478 -0.155  1.000
```

2. Plot the data

```r
par(mfrow = c(2, 2))
plot(anscombe$x1, anscombe$y1, main = "Dataset 1", pch = 19)
abline(lm(y1 ~ x1, data = anscombe), col = "red")

plot(anscombe$x2, anscombe$y2, main = "Dataset 2", pch = 19)
abline(lm(y2 ~ x2, data = anscombe), col = "red")

plot(anscombe$x3, anscombe$y3, main = "Dataset 3", pch = 19)
abline(lm(y3 ~ x3, data = anscombe), col = "red")

plot(anscombe$x4, anscombe$y4, main = "Dataset 4", pch = 19)
abline(lm(y4 ~ x4, data = anscombe), col = "red")
```