19 データ分析プログラミング (3)

花澤楓 学籍番号: 2125242

2023/11/20

1　図の作成

図表は目的（何を見せたいのか）に応じて様々な種類がある。目的は大きく分けて 4 つで、1. 比較、2. 分布、3. 関係性、4. 構成（Composition）がある。図の文法（grammar of graphics）に従い、データの各変数を**視覚的な要素**に対応づけることが求められる。よって、図表は各変数から視覚的情報への関数のようなもの。

視覚的な要素とは、以下の通り。

- 位置：x 軸、y 軸。座標軸（直交座標・デカルト座標、円形など）に「形」としてデータを配置するため
- 大きさ：強調したいデータは大きくするなど
- 形：各 group ごとにプロットの形（丸、三角、四角など）を変えるなど
- 色：寒色系と暖色系のよい組み合わせを選び、色弱・色盲の方への配慮も必要
- ラベルづけ

R には、`ggplot2` パッケージが用意されており、`base R` での `plot` よりも、より柔軟性がある関数でかつデフォルトの状態でプロットしても綺麗に見えやすいことからよく利用されている。また、特徴として図を段階的に重ねていくようなイメージで作成できる。まず、ベースとなる座標軸を用意し、データを重ね、ラインを引き、ラベル付けをする、といったイメージ。

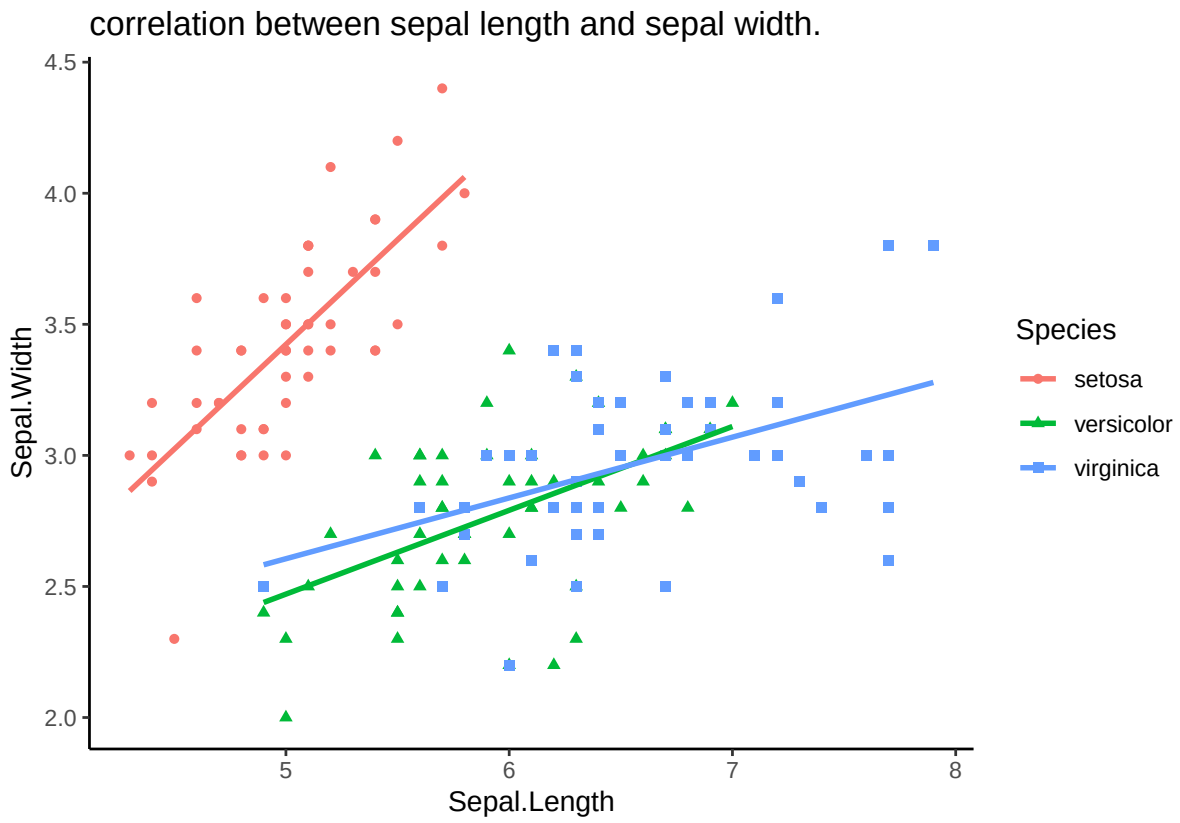
以下に、`iris` データを使用して `ggplot2` の使用例を示す。

```
plot_object <- ggplot(data = iris,
                      mapping = aes(x = Sepal.Length,
                                    y = Sepal.Width,
                                    color = Species,
                                    shape = Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_classic()
```

```
plot_object <- plot_object +
  labs(title = "correlation between sepal length and sepal width.")

plot_object
```

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



図を保存する際には、.png でなく.pdf で保存することで、数式として埋め込まれるので拡大しても荒くならない。そのため、論文やレポートに示す際には.pdf が推奨される。

また、同じデータから生成されるグラフでも、見せ方の違いによって図やグラフを見る人の受ける印象が全く異なる。

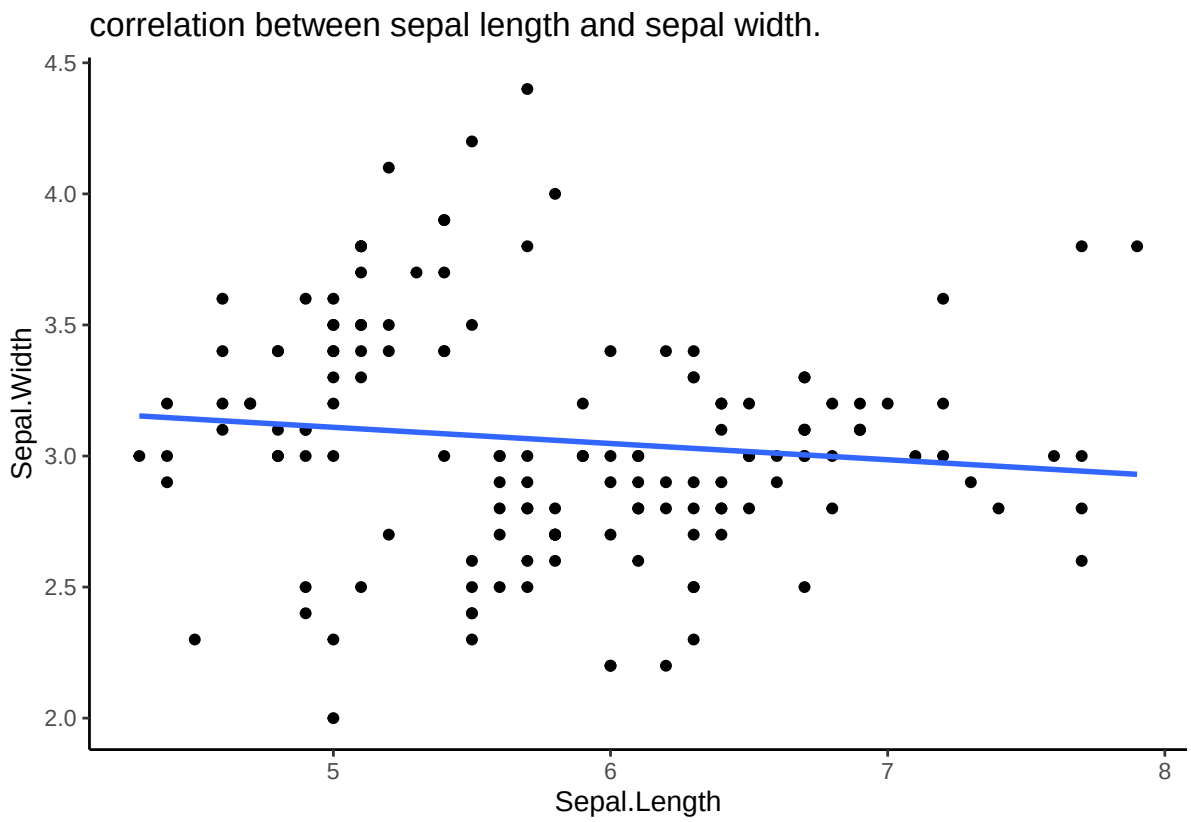
例えば、先ほどの `iris` データにおいて Sepal width と length の関係を見る際に、Species ごとに分類分けしないでプロット・回帰直線の表示をすると、

```
plot_object <- ggplot(data = iris,
                      mapping = aes(x = Sepal.Length,
                                    y = Sepal.Width)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
```

```
  theme_classic() +
  labs(title = "correlation between sepal length and sepal width.")

plot_object
```

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



となる。このグラフだけを見ると、ほとんど相関していないように見え、実際に相関係数を計算すると、

```
cor(iris$Sepal.Length, iris$Sepal.Width)
```

```
## [1] -0.1175698
```

負の相関係数となっている。相関係数を Species ごとに計算すると、

```
correlation_by_species <- iris %>%
  group_by(Species) %>%
  summarize(correlation_coefficient = cor(Sepal.Length, Sepal.Width))

correlation_by_species
```

```
## # A tibble: 3 x 2
##   Species    correlation_coefficient
##   <fct>                        <dbl>
## 1 setosa                       0.743
## 2 versicolor                   0.526
## 3 virginica                    0.457
```

と、いずれも正の相関が見られる。これは、先ほどの Species ごとに分類した上で回帰直線を引いているグラフを見れば明らか。このように、同じデータでも全く異なる見せ方ができてしまうので正しい倫理観のもと研究内容を報告すべきである。