# 19 データ分析プログラミング (2)

花澤楓 学籍番号: 2125242

2023/11/13

## 1　統計モデルの推定

統計モデルでは、データが持つ情報を回帰係数などの結論に集約し、モデルとデータが適合するような結論（パラメータ）の値を求める。何らかの目的変数を最適化する問題が多い。

**分析者の立ち位置**

統計モデルについて、手法の研究・開発をメインでする研究者と、開発された手法を適切に用いて実証・応用的な分析を行う研究者が存在する。手法開発者は、利用の利便性のために、R や Python で誰でも開発された手法を利用できるようなパッケージ・関数を作成する。また、分析実践者はすでに作成された関数を適切に使うことが求められる。統計モデルへの理解が不十分である場合、誤った解釈をしてしまうことがあるので注意が必要。

**推定のステップ**

1. データの入手・加工 + データの変数とモデルの変数の対応づけ
2. チューニングパラメータをインプット
3. モデル推定のための何らかの関数 `package::function()` を使用
4. 関数から出力された推定値のリストを取得。結果の図示（表・図に出力）

例えば、`iris` において、アヤメのがく片の長さ (Sepal.Length) をがく片の幅、花弁の長さ、花弁の幅から OLS 推定することを考える。モデル推定のために lm 関数を使用し、結果の表を `modelsummary::modelsummary()` より示す。

```
results_lm <- lm(Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width, data=iris)
modelsummary::modelsummary(results_lm, output = "markdown")
```

| | (1) |
|---|---|
| (Intercept) | 1.856 |
| | (0.251) |

| | (1) |
|---|---|
| Sepal.Width | 0.651 |
| | (0.067) |
| Petal.Length | 0.709 |
| | (0.057) |
| Petal.Width | -0.556 |
| | (0.128) |
| Num.Obs. | 150 |
| R2 | 0.859 |
| R2 Adj. | 0.856 |
| AIC | 84.6 |
| BIC | 99.7 |
| Log.Lik. | -37.321 |
| F | 295.539 |
| RMSE | 0.31 |

# 2 文書の作成

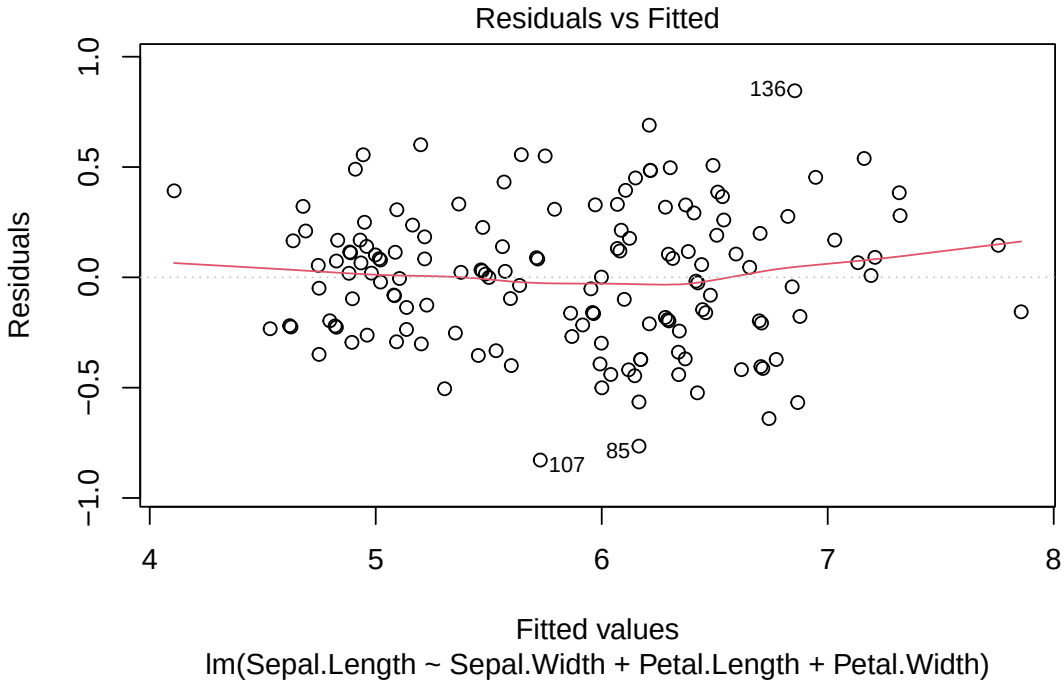
データ分析のレポートでは、文章が 75%、図表を 25% 程度にすることが一般的である。文章は、一次元のデータで、多くの情報量を含むことができるが視覚的にはわかりにくい場合が多い。図表にデータを起こす場合にはデータが二次元となり視覚的な情報を多く与えてくれる。そのため、データ分析の際には、raw データのみではなく推定結果を示す図表も同時に示すことが重要である。

## 2.1 データの視覚化

データを視覚化することには主に 2 つの目的がある。

1. 視覚的伝達
   データを視覚化することで、人間の認知機能が持つ視覚的思考を生かすことができる。また、計量経済学や因果推論に関する研究では、結果を 1 つのグラフにまとめることが可能なため、積極的に利用することが重要である。視覚化することで、文章のみではわからなかったような有意義な差異・比較をすることができる。例えば、先ほどの `iris` データから推定した結果の、残差と線形重回帰による予測値の２要素による図を生成する。縦軸が残差、横軸が予測値。

```
plot(results_lm, which = 1)
```




2. 探索的発見

   視覚化すると、推定値のみではわからないパターンを発見することができる。例えば、Residuals vs Fitted のグラフを見たときに、Residuals が非線形（山なりなど）に分布している場合、線形回帰モデルでは真のモデルを捉えきれていないことがわかる。