13 モデルの適合度と複雑性（１）

花澤楓 学籍番号: 2125242

2023/12/04

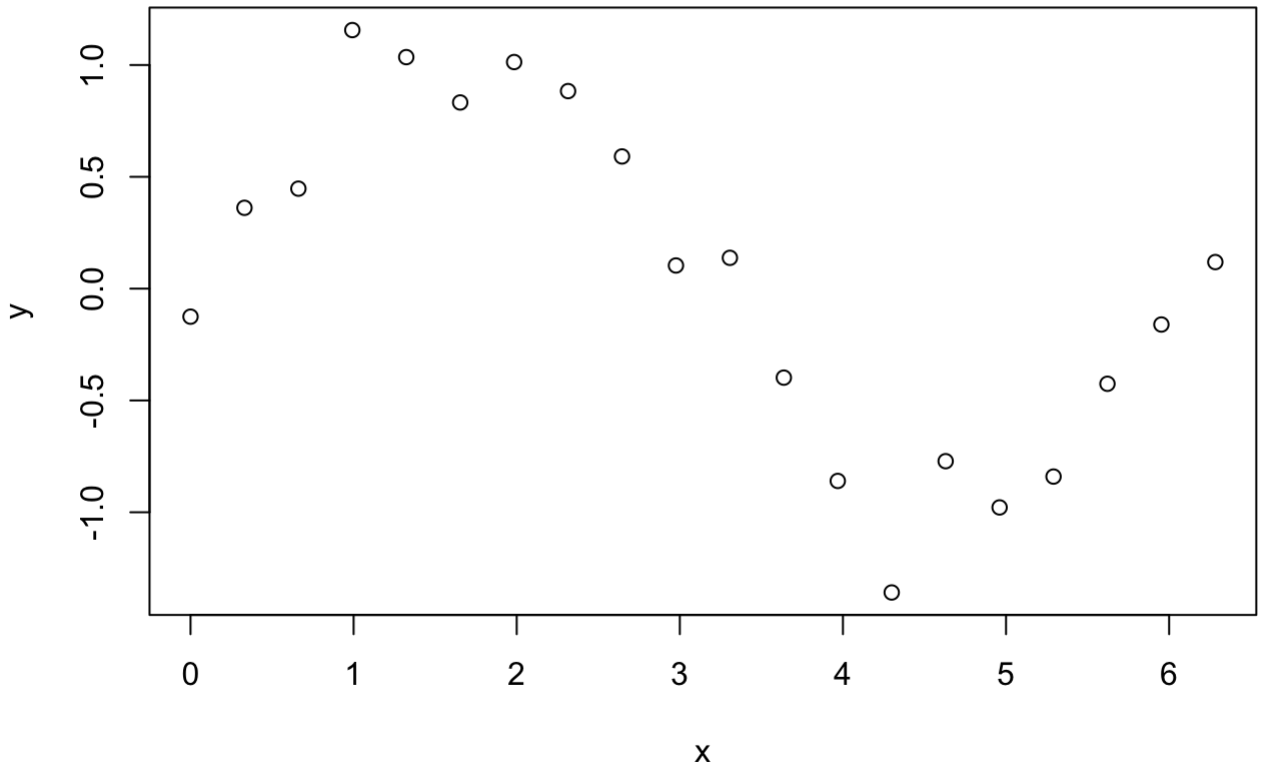
モデルの適合（fit）と複雑性（complexity）

2つの文化

情報通信産業の発達により分析できるデータの数が指数関数的に増加している。そのため、データに当てはまりが良いモデルを考える際に線形回帰を仮定する必要は必ずしもなく、非線形モデルを適合させた方がより当てはまりが良い場合が多い。例えば、以下のようなデータを考える。

```
# sinxにノイズを追加したデータセットの生成
# サンプル数
n <- 20

set.seed(1)
x <- seq(0, 2*pi, length.out = n)
y <- sin(x) + rnorm(n, sd = 0.2)
plot(x, y)
```



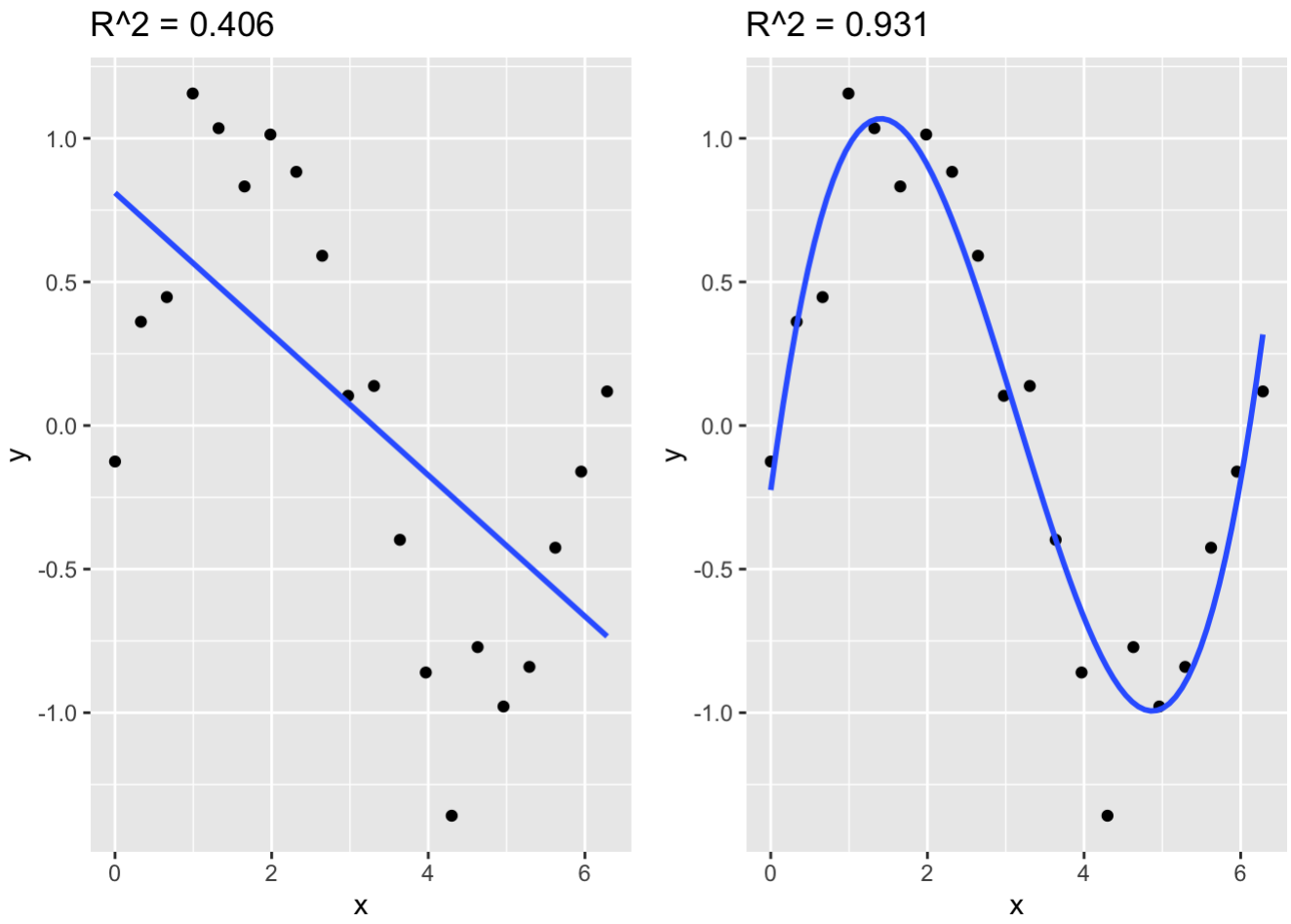
このとき、線形回帰モデルと非線形回帰モデルをそれぞれを当てはめると以下のようになる。

```
df <- data.frame(x, y)
# 線形回帰モデルによって当てはめ
OLS_model <- lm(y ~ x, data = df)
OLS_r2 <- summary(OLS_model)$r.squared
OLS <- df |>
  ggplot(aes(x, y)) +
  geom_point() +
  geom_smooth(method = "lm", se=FALSE) +
  ggtitle(paste0("R^2 = ", round(OLS_r2, 3)))

poly_model <- lm(y ~ poly(x, 3), data = df)
poly_r2 <- summary(poly_model)$r.squared
# 多項式(3次)モデルによって当てはめ
nonlinear <- df |>
  ggplot(aes(x, y)) +
  geom_point() +
  geom_smooth(method = "lm",formula = y ~ poly(x, 3), se=FALSE) +
  ggtitle(paste0("R^2 = ", round(poly_r2, 3)))

# 2つの図を並べて表示
grid.arrange(OLS, nonlinear, ncol=2)
```

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



グラフより、明らかにモデルの適合度がよくなっていることがわかる。

現実に観測されるデータ$\{X_i, Y_i\}$は何らかのデータ生成モデル（線形回帰なら、$Y_i = \beta X_i + \epsilon$など）に従って、データが生成されていると考えられる。このとき、データ生成モデルの複雑性が高いほど、データに当てはまりが良くなるが、未知のデータに対する予測性能は低くなる。このような現象を**過学習（overfitting）**と呼ぶ。過学習を防ぐためには、データ生成モデルの複雑性とデータに当てはまりの良さのバランスを考える必要がある。また、伝統的な計量経済学では、モデル当てはまりを良くすることが目的ではなく、パラメータの推定を目的としている。データ分析において以下の2つの文化がある。

1. 科学理想主義・不偏性

   この考え方は学術研究におけるものであり、経済学ではパラメータの不偏性や一致性、分散や平均に関心がある。例えば、古典的なOLSではいくつかの仮定を置いてその下でのOLS推定量はBLUEであることを示している（ガウス・マルコフの定理）。つまり、$\hat{\beta}$に興味がある。

2. 実践主義・R^2・適合度・予測/アルゴリズム

   この考え方は、学術研究でもあるがより実務で必要とされているものである。実務上では手元のデータから何か予測をしたいことが多くある。そのような場合にはパラメータの性質を気にすることはなく、訓練データで学習したモデルがテストデータでどれだけ当てはまりがいいかに大きな関心がある。つまり、$\hat{y}$。

経済学での変遷

1. ノン・パラメトリック（non-parametric）推定
   - 誤った仮定（misspecification）を避けるため、データ生成モデルに特定の分布を仮定しない
   - モデルの複雑性がサンプル数に応じて（ゆっくりと）増加する
2. 機械学習（machine learning）
   - ビックデータ（画像、音声、テキストなど、デジタル・フットプリント）：3Vs（velocity, volume, variety）を使用することで精度の高いモデルの学習が可能になった
   - 人工知能（AI）：何が出力として正解かの明示的な答えを与えられずに、自ら学習していくモデル

モデル選択：モデルの「複雑性」を上げれば当てはまりは良くなるが、過学習などの問題から「精度」が落ちる。そのため、これら2つの間にはトレードオフが存在する。そのため、必要最低限の仮定でモデルを立てることが必要（オッカムの剃刀）。また、「複雑性」が高くなるほど出力された結果の解釈が難しくなる。

モデル集計：データから複数のモデルを学習させ、最も良い精度のモデルを採用する（多数決や平均などを計算）アンサンブル法（ensemble method）がある。例えば、ランダムフォレスト（random forest）や勾配ブースティング（gradient boosting）などがある。

適合度の分解と「複雑性」の最適化

データ$\{X_i, Y_i\}$の生成モデルを以下のように仮定する。

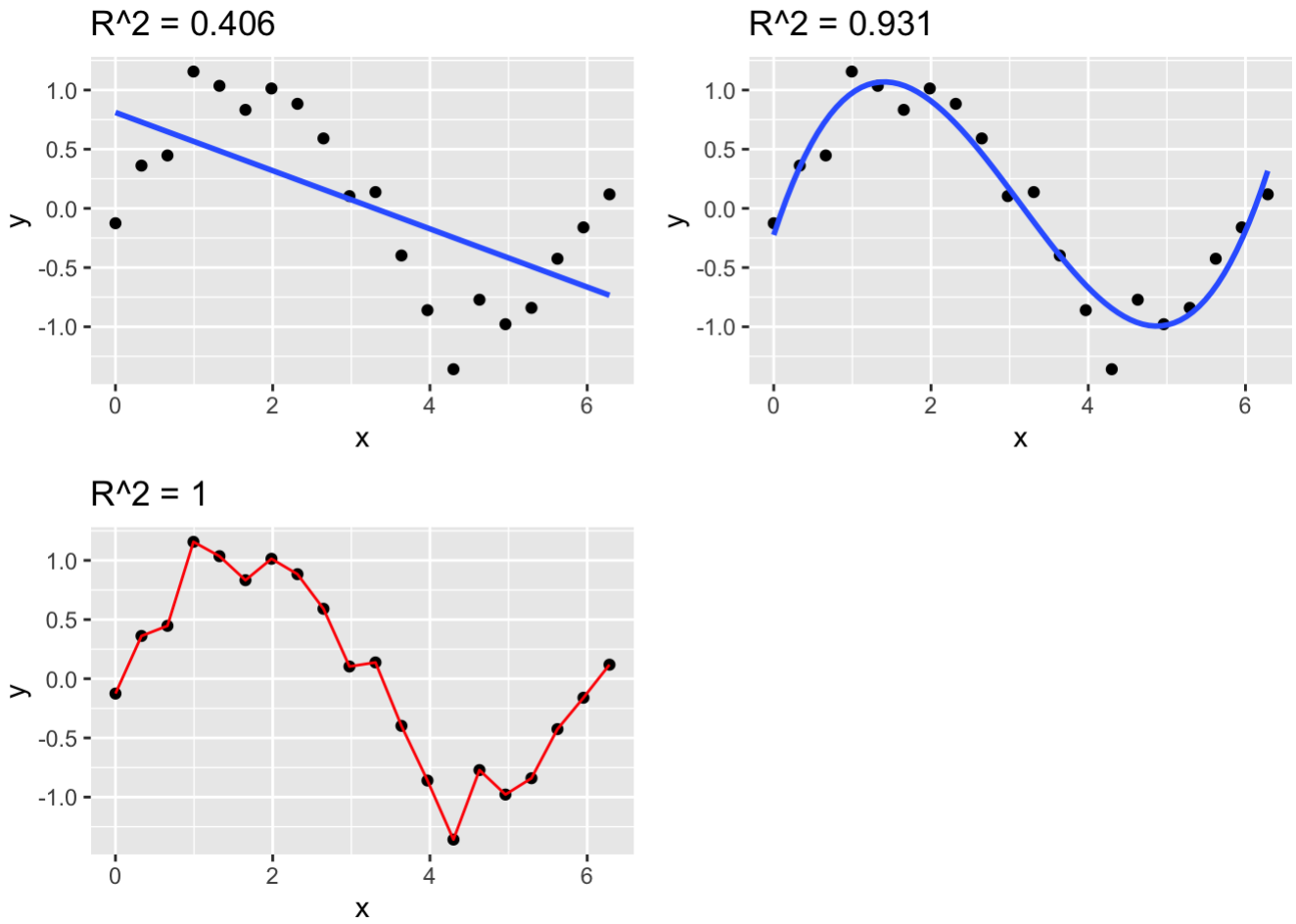
$$Y_i = f(X_i) + \epsilon$$

ここで、$f(X_i)$は未知の関数であり、ϵは誤差項である。このとき、$f(X_i)$を推定することが目的となる。例えば、$f(X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_k + \epsilon$と仮定する。ここで、$k$はサンプル数。このモデルを先ほど生成したデータに当てはめてみると以下の通り。

```
# n次多項式モデル
poly_full <- lm(y ~ poly(x, n-1), data = df)
poly_full_r2 <- summary(poly_full)$r.squared
poly_full_plot <- ggplot(df, aes(x, y)) +
  geom_point() +
  geom_line(aes(y = predict(poly_full)), color = "red") +
  ggtitle(paste0("R^2 = ", round(poly_full_r2, 3)))

grid.arrange(OLS, nonlinear, poly_full_plot, ncol=2)
```

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



これは、手元のデータについてyをxで完璧に説明できていることを示す。しかし、これは過学習が起きていて同じデータ生成過程から生じる別のデータセットに対しての当てはまりは良くない。

予測誤差のバイアス・バリアンス分解：予測誤差は以下のように分解できる。

$$
\begin{aligned}
\text{予測誤差} &= E(Y_i - \hat{f}(X_i))^2 \\
&= E(f(X_i) + \varepsilon - \hat{f}(X_i))^2 \\
&= E(f(X_i) - \hat{f}(X_i))^2 + E(\varepsilon^2) + 2E(f(X_i) - \hat{f}(X_i))E(\varepsilon) \\
&= E(f(X_i) - \hat{f}(X_i))^2 + V(\varepsilon) \\
&= E(f(X_i) - E(\hat{f}(X_i)^2)) + E(\hat{f} - E(\hat{f}(X_i)))^2 + 2E[f(X_i) - E(\hat{f}(X_i))] + V(\varepsilon) \\
&= \text{バイアス}^2 + \text{バリアンス} + \text{ノイズ}
\end{aligned}
$$

既知のデータ（訓練データ）に対してはモデルの複雑さが増すほどerrorは少なくなるが、未知のデータに対してはバリアンスが大きくなってしまうトレードオフが存在する。既知のデータに対して当てはまりをよくすれば（バイアスを減らす）するほど、未知のデータに対して当てはまりが悪くなる過学習が生じるため、損失関数にバイアスを入れることで適合度を改善する、正則化といった技術がある。