第3回：非線形回帰

Code

1. 前準備

これまで解説した線形回帰は、目的変数を説明変数の1次式で表すような統計モデルのことでした。しかし、一般には目的変数と説明変数の間の関係が1次式でない（非線形である）ような場合も考えられます。このような場合に用いられる統計モデルが非線形回帰（nonlinear regression）です。今回は、非線形回帰にまつわる話題から、

- 非線形回帰の概要
- 多項式回帰
- べき乗回帰
- モデル選択

を解説します。

デモデータには、 R 言語にデフォルトで読み込まれている mtcars データセットを用います。このデータセットは、自動車の馬力と燃費の関係を記録したものです。

Hide

```
# mtcarsデータセットの確認
dat <- mtcars[, c("hp", "mpg")]
head(x = dat, n = 5)
```

	hp <dbl>	mpg <dbl>
Mazda RX4	110	21.0
Mazda RX4 Wag	110	21.0
Datsun 710	93	22.8
Hornet 4 Drive	110	21.4
Hornet Sportabout	175	18.7

5 rows

標本サイズと次元、変数の名前を確認しておきましょう。

Hide

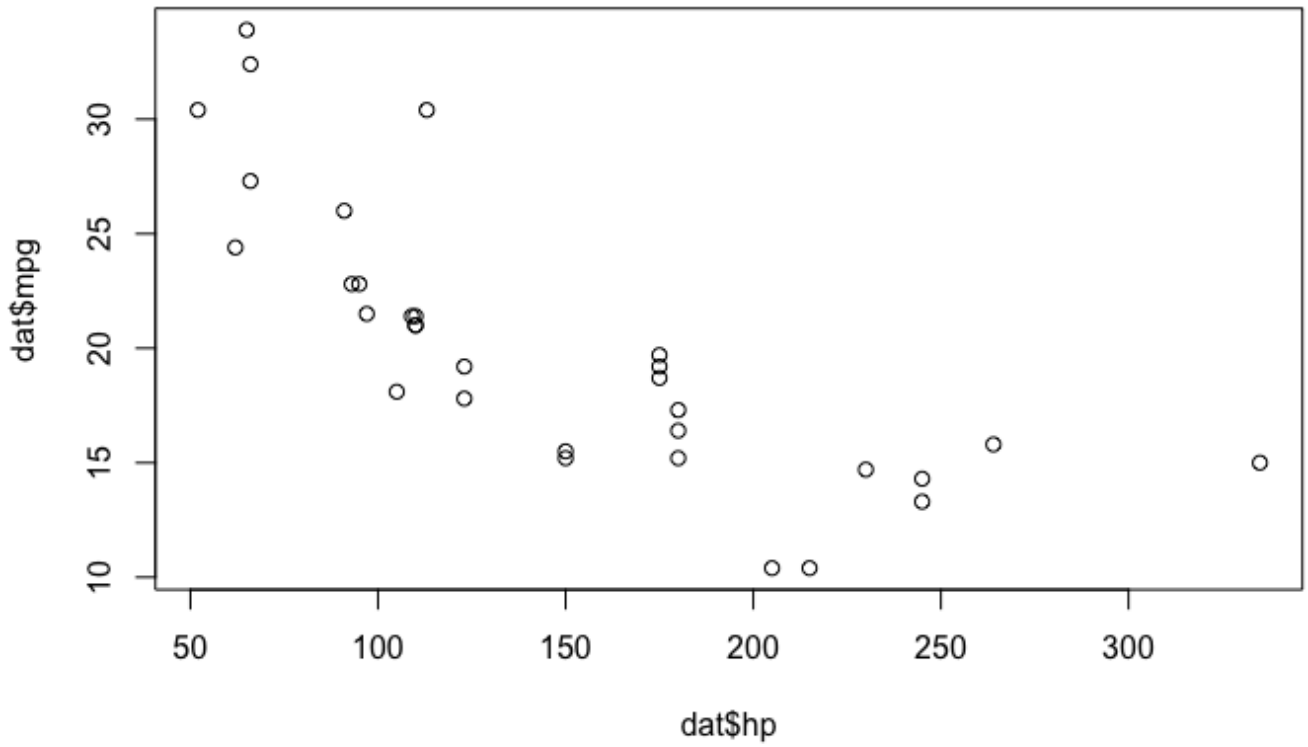
```
# 標本サイズ、次元、変数の名前の確認
str(dat)
```

```
'data.frame':	32 obs. of  2 variables:
 $ hp : num  110 110 93 110 175 105 245 62 95 123 ...
 $ mpg: num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

今回は、馬力 hp を用いて燃費 mpg を説明するモデルを作ろうと思います。2変数の関係を把握するために散布図をかいておくことで、 馬力 と 燃費 との間は非線形な関係にあることが確認できます。

Hide

```
# 散布図
plot(dat$hp, dat$mpg)
```



2. 非線形回帰

この節では、非線形回帰の概要を説明します。また、非線形回帰の具体例として、

- 多項式回帰
- べき乗回帰

を紹介します。

2.1 非線形回帰の概要

A. 課題設定

データに D 個の変数 $x_1, \cdots, x_D$ と y が記録されているとき、変数 y を変数 $x_1, \cdots, x_D$ の1次式ではない（非線形な）関数

$$y = f(x_1, \cdots, x_D; \beta) + \text{誤差}$$

で表現することで説明・予測できると仮定します。例えば、燃費を馬力で説明する式として

$$\text{燃費} = \beta_0 + \beta_1 \times \text{馬力} + \beta_2 \times \text{馬力}^2 + +\text{誤差}$$

という2次式や、

$$\text{燃費} = \beta_1 \times \text{馬力}^{\beta_2} + \text{誤差}$$

という式で説明できないかを検討していきましょう。前者のように説明変数の多項式で目的変数のとりうる値の傾向を表現するようなモデルを**多項式回帰**（polynomial regression）、後者のように説明変数の累乗とその定数倍で目的変数のとりうる値の傾向を表現するようなモデルを**べき乗回帰**（power regression）といいます。

B. 出来ること

モデルに含まれる値が未知のパラメータ $\beta_1, \cdots, \beta_p$ を**偏回帰係数**（coefficient）といいます。これらの係数を推定することで、各説明変数の値が決まっているとき、目的変数の値はいくらと予測できるかを推定することができます。また、偏回帰係数の区間推定や統計的仮説検定を行うことができます。これについてはデモで例を交えながら解説します。

C. 偏回帰係数の推定の仕組み

偏回帰係数の推定値 $\hat{\beta}$ は、データ全体で

- 偏回帰係数から決まる各標本点の予測値 $\hat{y} = f(x_1, \cdots, x_D; \hat{\beta})$
- 実測値 y

の差（**残差**といいます）の2乗 $(y - \hat{y})^2$ の平均値

$$l(\beta_0, \beta_1, \cdots, \beta_D) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

が最小になるように決めます。これを**最小2乗法**（least squares method）といいます。非線形回帰とは、目的変数と説明変数の関係に偏回帰係数 β で特徴付けられる関数 $f(x_1, \cdots, x_D; \beta)$ を仮定したとき、これを最小2乗法で推定する手法のことです。

2.2 多項式回帰のデモ

燃費 mpg を馬力 hp の2次式で表現するモデル、つまり

$$\text{燃費} = \beta_0 + \beta_1 \times \text{馬力} + \beta_2 \times \text{馬力}^2 + \text{誤差}$$

を考えます。多項式回帰では、説明変数の値をそのまま用いると、大きい次数の項の値がとても大きくなったり小さくなったりすることで、計算機が数値的に不安定になることがあります。そのため、説明変数を事前に正規化（例えば標準化やmin-max正規化）することが一般的です。今回は、min-max正規化を採用します。

Hide

```
# `hp`のmin-max正規化を`hp_scale`という名前で作る。
min_hp <- min(dat$hp)
max_hp <- max(dat$hp)
dat$hp_scale <- (dat$hp-min_hp) / (max_hp-min_hp)
head(x = dat, n = 5)
```

	hp <dbl>	mpg <dbl>	hp_scale <dbl>
Mazda RX4	110	21.0	0.2049470
Mazda RX4 Wag	110	21.0	0.2049470
Datsun 710	93	22.8	0.1448763
Hornet 4 Drive	110	21.4	0.2049470

	hp <dbl>	mpg <dbl>	hp_scale <dbl>
Hornet Sportabout	175	18.7	0.4346290

5 rows

多項式回帰の場合、 R 言語では次のように偏回帰係数を計算することができます。

Hide

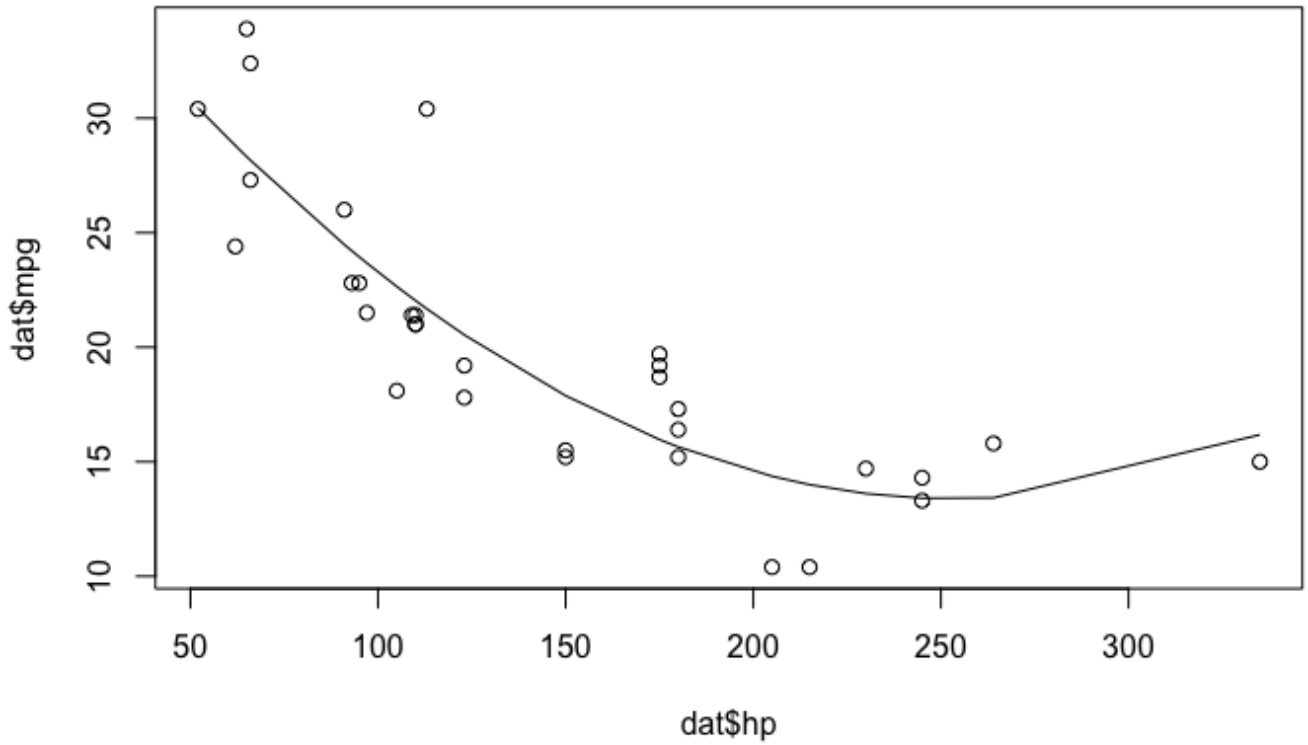
```
# 多項式回帰
result_deg2 <- lm(mpg ~ poly(hp_scale, degree = 2, raw = TRUE),
                  data = dat)
summary(result_deg2)
```

```
Call:
lm(formula = mpg ~ poly(hp_scale, degree = 2, raw = TRUE), data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5512 -1.6027 -0.6977  1.5509  8.7213 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                           30.455      1.309  23.271  < 2e-16 ***
poly(hp_scale, degree = 2, raw = TRUE)1  -47.981      7.085  -6.772 1.96e-07 ***
poly(hp_scale, degree = 2, raw = TRUE)2   33.703      7.884   4.275 0.000189 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.077 on 29 degrees of freedom
Multiple R-squared:  0.7561,    Adjusted R-squared:  0.7393 
F-statistic: 44.95 on 2 and 29 DF,  p-value: 1.301e-09
```

また、得られた2次関数を散布図上に重ねてかくと、次のようになります。

Hide

```
plot(dat$hp, dat$mpg)
lines(sort(dat$hp), predict(result_deg2)[order(dat$hp)])
```



問題 : 以下の問いに答えてください。

1. 推定によって得られた2次式を答えてください。
2. 推定された結果に不自然な点があるかを検討してください。

解答 :

1. 得られた3次式は以下の通りです。なお、式の中に現れる「馬力」は、min-max正規化を施した後の変数とします。

$$\text{燃費} = -30.45 + 2.38 \times \text{馬力} + 0.08 \times \text{馬力}^2 + \text{誤差}$$

2. 馬力が大きいほど燃費は必ず悪くなるとなると想定されます。しかしグラフからは、馬力が 250 より大きくなると燃費が悪くなるという予測を与える式になっていることが確認できます。これが、このモデルの不自然な点です。■

2.3 べき乗回帰のデモ

燃費 mpg を馬力 hp の3次式で表現するモデル、つまり

$$\text{燃費} = \beta_1 \times \text{馬力}^{\beta_2} + \text{誤差}$$

を考えます。べき乗回帰では、偏回帰係数を推定するために、事前に初期値を決めておく必要があります。べき乗回帰における最小2乗法は**非線形最小2乗法**といって、予め決めた偏回帰係数の初期値から、より残差の2乗和が小さくなるような偏回帰係数の値へと更新していくような計算方法で、偏回帰係数の値を求めています。

この初期値を求める方法の一つとして、次のような線形回帰を先に計算しておきます。

Hide

```
# 偏回帰係数の初期値の決定
lm(log(mpg) ~ log(hp), data = dat)
```

```
Call:
lm(formula = log(mpg) ~ log(hp), data = dat)

Coefficients:
(Intercept)      log(hp)
     5.5454      -0.5301
```

問題 : 以下の問いに答えてください。

1. べき乗回帰の式 $\text{燃費} = \beta_1 \times \text{馬力}^{\beta_2} + \text{誤差}$ の「誤差」を無視したとき、両辺の対数をとって得られる式を答えてください。
2. 上の線形回帰の結果から、偏回帰係数の初期値をどう決めると良いかを考えてみてください。

解答 :

1. $\log \text{燃費} = \log \beta_1 + \beta_2 \log \text{馬力}$
2. Intercept は5.5454、 log(hp) の偏回帰係数は-0.5301でした。(1)の結果から $\log \beta_1 = 5.5454$ とおいたとき $\beta_1 = \exp(5.5454)$、$\beta_2 = -0.5301$ と決めることができます。■

Remark : なお目的変数を $\log$ 燃費、説明変数を $\log$ 馬力 とした線形回帰を式で表すと、
$\log \text{燃費} = \alpha_1 + \alpha_2 \log \text{馬力} + \text{誤差}$ になります。指数をとると

$$\begin{aligned} \text{燃費} &= \exp \alpha_1 \times \text{馬力}^{\alpha_2} \times \exp(\text{誤差}) \\ &= \beta_1 \times \text{馬力}^{\beta_2} \times \exp(\text{誤差}) \qquad \beta_1 = \exp \alpha_1, \beta_2 = \alpha_2 \text{とおく。} \end{aligned}$$

になるので、誤差のモデリングがべき乗回帰とは異なることが確認できます。つまり、 lm 関数で得られた推定結果はべき乗回帰そのものの結果ではなく、べき乗回帰は改めてやり直す必要があります。■

得られた初期値を用いて、べき乗回帰を計算してみましょう。 R 言語では nls 関数を用いて、べき乗回帰の偏回帰係数を計算することができます。

Hide

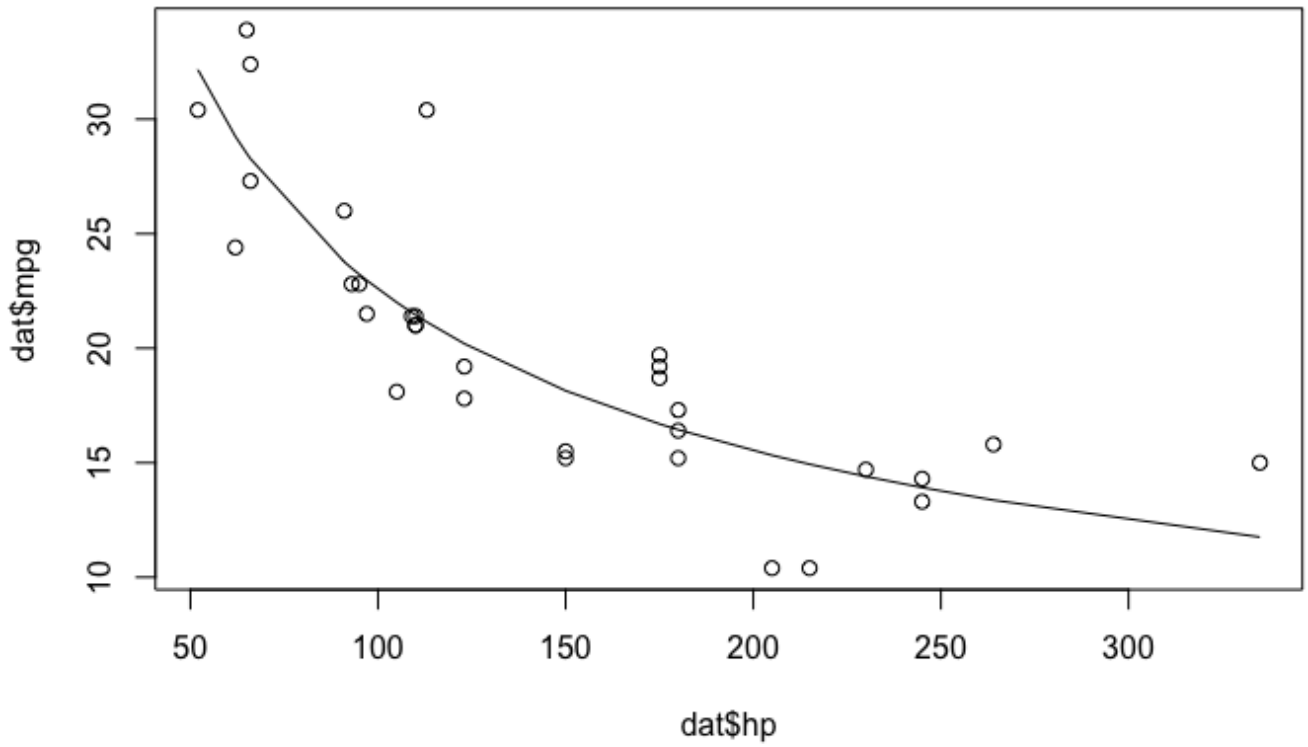
```
# 多項式回帰
result_power <- nls(mpg ~ a*hp^b,
                    data = dat,
                    start = list(a = exp(5.5454), b = -0.5301),
                    control = nls.control(maxiter = 100))
summary(result_power)
```

```
Formula: mpg ~ a * hp^b

Parameters:
   Estimate Std. Error t value Pr(>|t|)
a 272.11754   74.03692   3.675 0.000924 ***
b  -0.54043    0.05826  -9.277 2.55e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.102 on 30 degrees of freedom

Number of iterations to convergence: 2
Achieved convergence tolerance: 3.411e-06
```

また、得られた関数を散布図上に重ねてかくと、次のようになります。

Hide

```
plot(dat$hp, dat$mpg)
lines(sort(dat$hp), predict(result_power)[order(dat$hp)])
```



問題 : 推定によって得られた関数を答えてください。

解答 :

得られた関数は以下の通りです。

$$\text{燃費} = 272.12 \times \text{馬力}^{-0.54} + \text{誤差}$$

2.3 モデル選択

A. 赤池情報量規準

今回は、燃費 mpg を馬力 hp で表現する式を、多項式回帰とべき乗回帰の2通りで作りました。複数のモデルを作ったら、どちらのモデルを採用するかを検討しましょう。多項式回帰の節で出題した問題のように、各モデルの長所・短所を検討してモデルを選択することが大切です。しかし、場合によっては、それでもモデル選択に悩むことがあるでしょう。このようなとき、モデル選択の際にヒントになる指標の一つとして、未知のデータへの予測の精度を推定する方法だった赤池情報量規準を用いることができます。（第2回資料も参考にしてみてください。）

Hide

```
# 多項式回帰とべき乗回帰の比較
AIC(result_deg2); AIC(result_power)
```

```
[1] 167.6023
[1] 167.1958
```

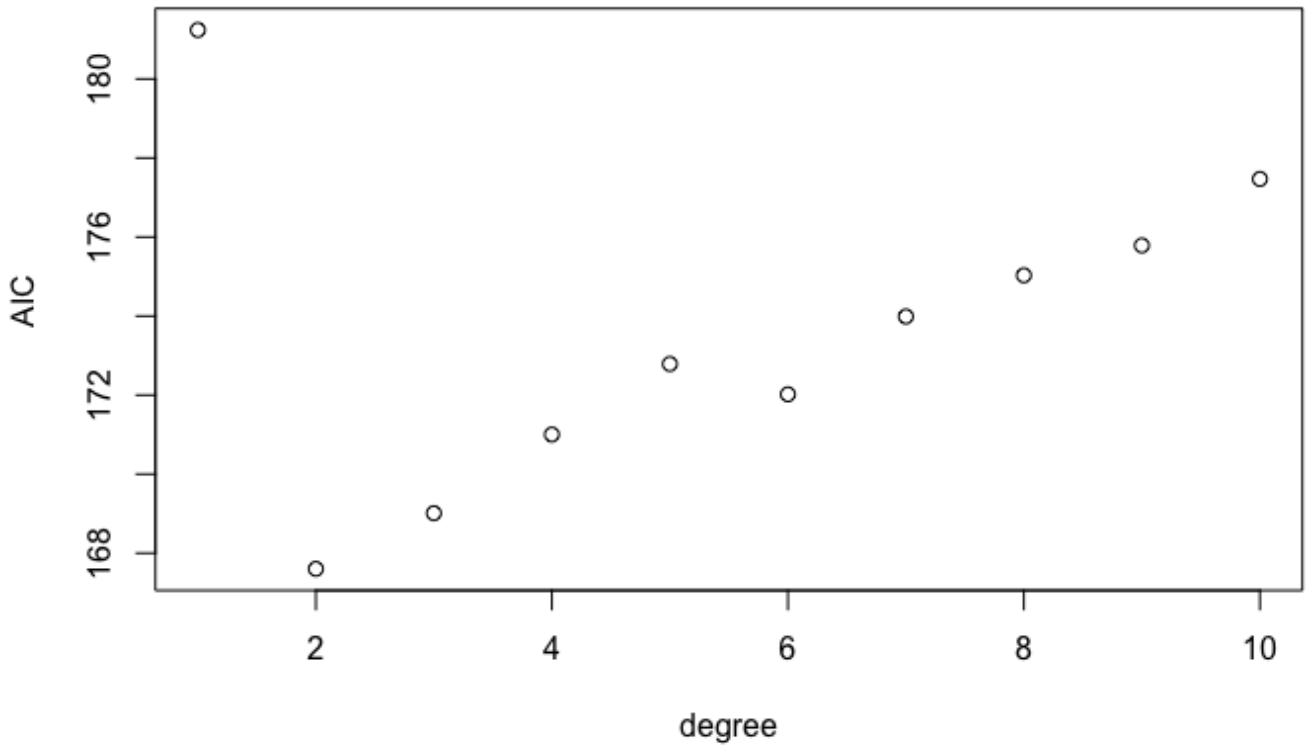
今回は、赤池情報量規準がより小さいべき乗回帰を採用するほうが良いのではないかと推定できます。

B. 多項式回帰の次数の決定と過剰適合

多項式回帰の次数を決定する際、以下のように赤池情報量規準を各次数で計算し図にまとめることがあります。例えば、以下の図からは次数が2の場合が最も赤池情報量規準が小さいことが確認できます。

Hide

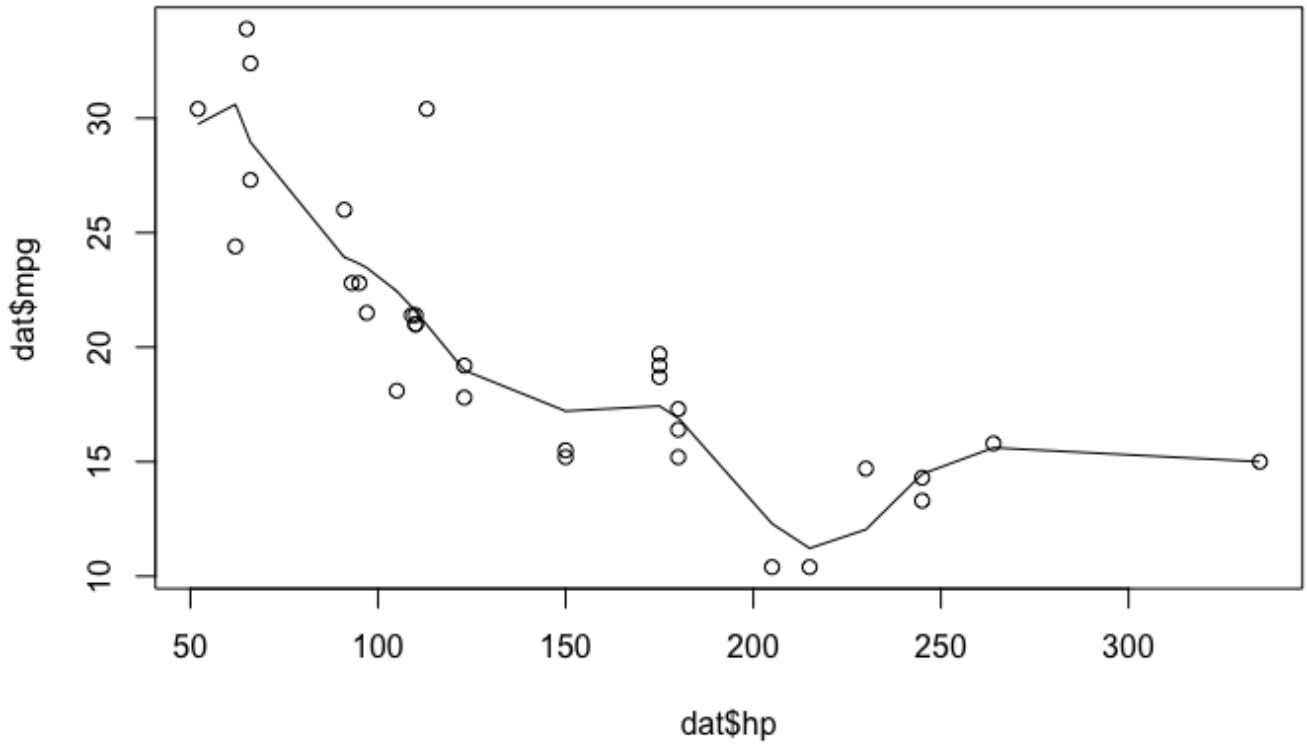
```
# 多項式回帰の次数の決定
aic <- c()
for (d in 1:10) {    # 次数1から10までの多項式回帰を逐次計算する。
  result_poly <- lm(mpg ~ poly(hp_scale, degree = d, raw = TRUE),
                    data = dat)
  aic[d] <- AIC(result_poly)
}

plot(1:10, aic, xlab = "degree", ylab = "AIC")
```



ところで、試しに次数10の多項式回帰を実行すると、以下のようにデータによくあてはまっているように見えます。それにも関わらず赤池情報量規準が次数2の場合より高い値になっているのはなぜでしょうか。

Hide

```
# 次数10の多項式回帰（10次回帰ともいいます。）
result_deg10 <- lm(mpg ~ poly(hp_scale, degree = 10, raw = TRUE),
                   data = dat)
plot(dat$hp, dat$mpg)
lines(sort(dat$hp), predict(result_deg10)[order(dat$hp)])
```



理由は、赤池情報量規準が「未知のデータへの予測の精度」を推定する方法だからです。手元のデータによくあてはまることと、まだ得られていない未知のデータによくあてはまることは異なります。前者が起こって後者が起こらないような現象を**過剰適合**（過学習, overfitting）、その逆を**過少適合**（未学習, underfitting）といいます。これらは特に、続・初級統計学で扱う後半の話題で、とても重要な話になっていきます。