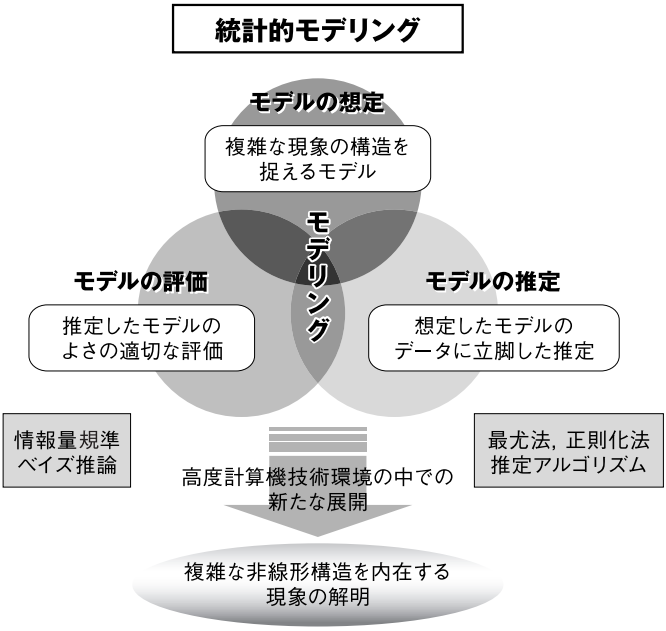
1
統計的モデリング

小西貞則

21 世紀の高度情報技術環境のなか，生命科学，情報技術，地球環境科学，システム工学，金融工学など，さまざまな科学技術分野において，それらはわれわれがかつて予測しえなかった勢いで大きな発展を遂げ，新たな学問分野の創始・創出が続いている．このような状況のなかで，さまざまな現象の情報源であるデータに基づく情報抽出と知識発見、複雑な自然現象や社会現象の解明とその予測・制御にあたっては，対象となる現象のモデリングが不可欠である．

ここでは，回帰，判別法における線形，非線形モデリングの基礎的な概念を述べた．今後も科学の諸分野から多様な問題が投げかけられ，現象分析に有効に機能する統計的データ解析手法の開発研究が常に必要と考えられる．本書が問題解決の一助になれば幸いである．



第 1 章

線形回帰モデル

データの背後にある現象の解明において，基礎的な役割を担うのが現象のモデル化である．本章では，現象の結果とそれに影響を及ぼすと考えられる複数の要因を結びつける最も基本的なモデルである線形回帰モデルについて述べる．線形回帰モデルを通して，モデルの想定，想定したモデルの推定，そして推定したモデルの評価という回帰モデリングの基本的な考え方について述べる．

1.1 2 変数間の関係を捉える

1.1.1 データとモデル

バネに力を加えると変形し，弾性限界内においてこの変形の大きさは加えた力に比例することは，フックの法則または弾性の法則としてよく知られている．図 1.1（左）は，バネに加えた力 (x g) とバネの長さ (y cm) を測定した 10 組のデータをプロットしたものである．

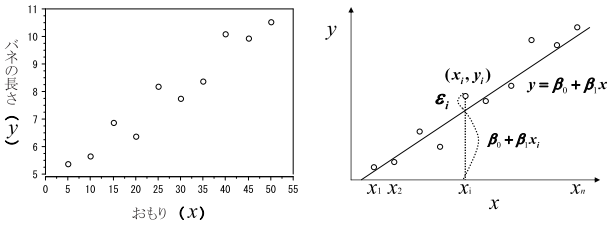


図 1.1 応力とバネの長さの測定データ (左) と線形回帰モデル (右)

測定の誤差を完全に制御できれば，この場合データは直線上に並ぶと考えられる．しかし，実際の観測データは誤差あるいはノイズを含んでおり，プロットすると図 1.1 のように直線的な傾向を示唆するものとなった．このように，

2
情報理論と学習理論

竹内純一

情報理論と学習理論は，ともにデータを扱うための理論である．前者は主に通信を目的とし，後者はデータから情報を獲得することを目的とする．本編の主役は，シャノン (Shannon) が定義した「情報量」である．ここでは，データ圧縮と機械学習における情報の数理を浮き彫りにすることを目指す．読者には，「学習はデータ圧縮である」という見方を身につけてくださることを期待している．

第1章

情報源符号化

大量のデータを，効率よく流通させたり，保存したりするには，そこから余分なものを取り除き，スリムにしておくことが望ましい．そのための技術がデータ圧縮である．

まず，どうすればデータをスリム化できるのか，例をあげて考えてみよう．以下のデータを考える．

00000000001111111111

これを

$0^5 1^5$

とすれば，ずいぶん小さくなる．これは，データ中にある繰り返しのパターンを利用して圧縮した例である．

また，次のデータを考えてみよう．

00000001000000000000000000001000000000000

これは，長さが 40 の文字列であり，8 番目と 28 番目が 1 であることを除き，ほとんど 0 である．こうしたデータは，珍しい文字である 1 のみに注目し，たとえば，

(40,8,28)

のように，列の長さと 1 の位置だけを記録する約束をすれば，小さなデータに書き換えることができる．これは，ありふれた文字は簡単に，珍しい文字は詳細に記述することで圧縮した例である．

ところで，これらの例では，使用する文字について注意を払わなくてはならない．もとのデータは ‘0’ と ‘1’ のみでできているが，圧縮後は 0〜9 の数字に加え，括弧や，べき乗を示す小さな右肩の数字を用いている．同じ長さの文字