カラー図版集

もともとモノクロの図版も含まれています.

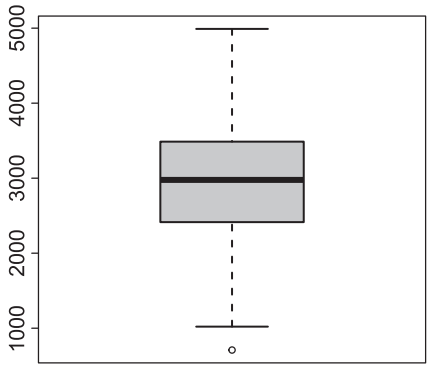

図 1.1　出生体重の箱ひげ図

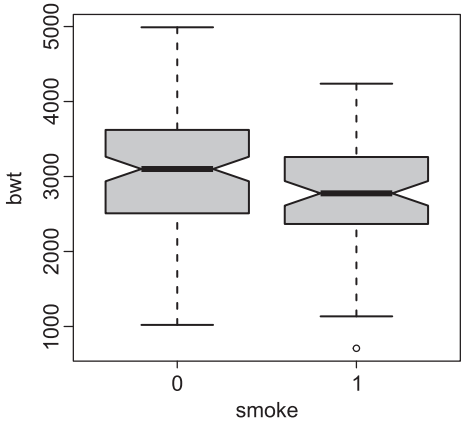

図 1.2　喫煙群別の出生体重の箱ひげ図

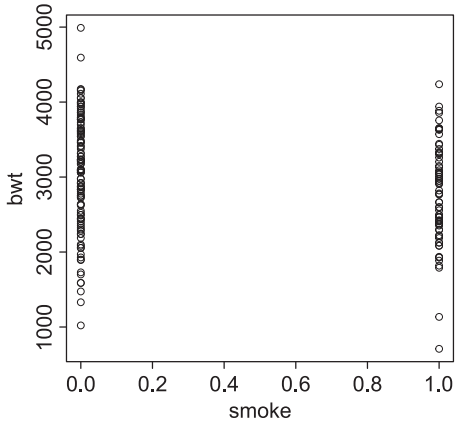

図 1.3　喫煙群別と出生体重の散布図

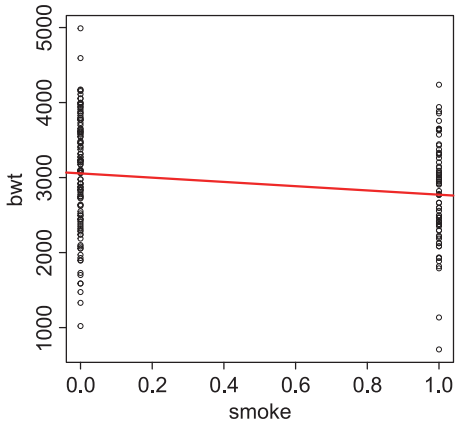

図 1.4　喫煙群別と出生体重の散布図に対する回帰直線

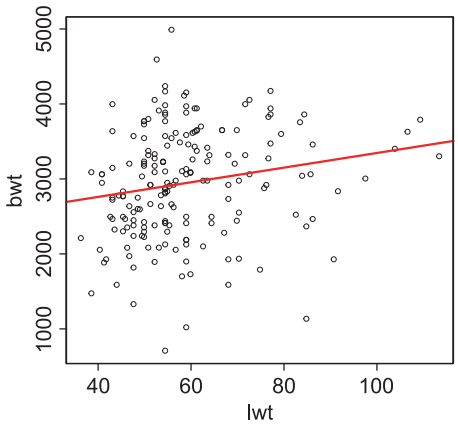

図 1.5　母の妊娠前の体重と出生体重の散布図に対する回帰直線

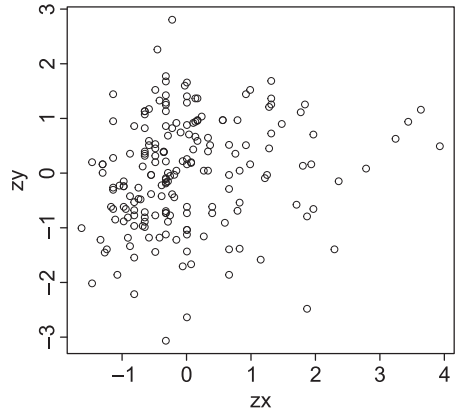

図 1.6　母の体重と出生体重の標準化データによる散布図

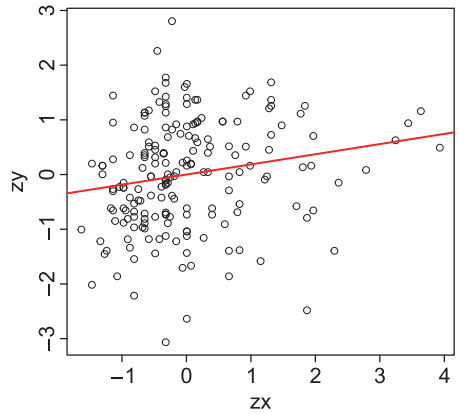

図 1.7　標準化データに対する回帰直線．回帰直線の傾きは標準化する前の元データに対する相関係数 0.186 に等しい．

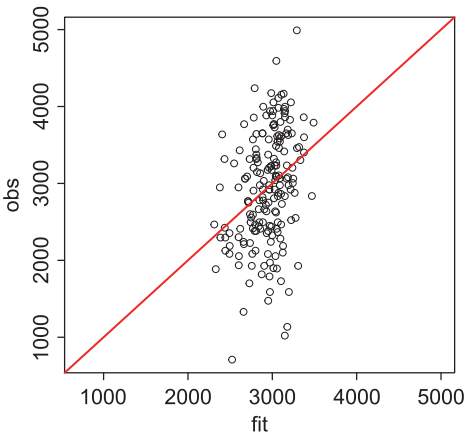

図 1.8　予測値と観測値の散布図と直線 $y = x$

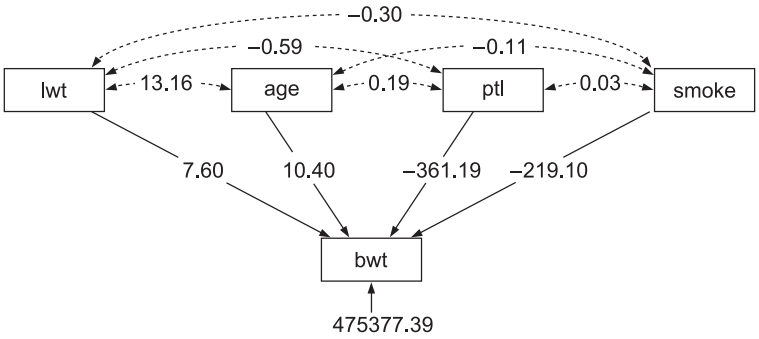

図 1.9　重回帰モデルのパス図による視覚化

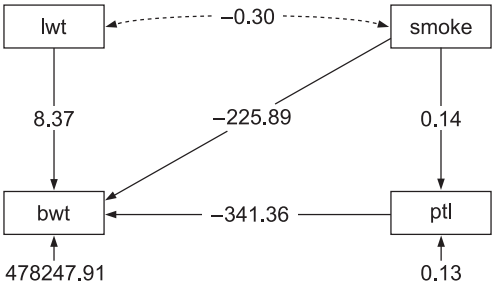

図 1.10　パス解析の視覚化

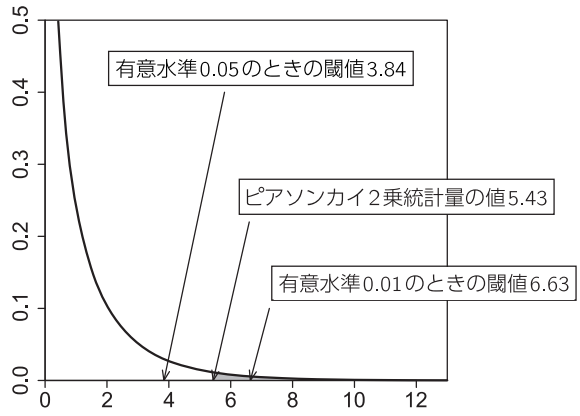

図 2.1　自由度 1 のカイ 2 乗分布の密度関数，各有意水準での閾値

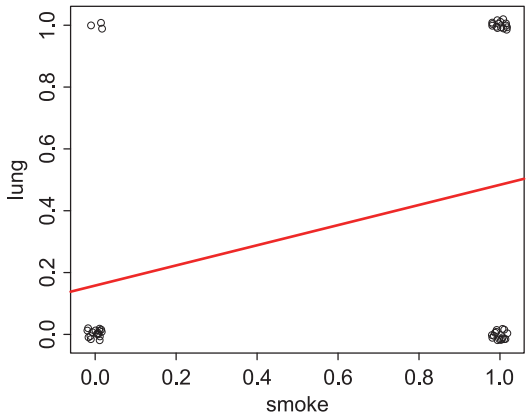

図 2.2　2 × 2 分割表データの散布図と線形回帰直線のあてはめ

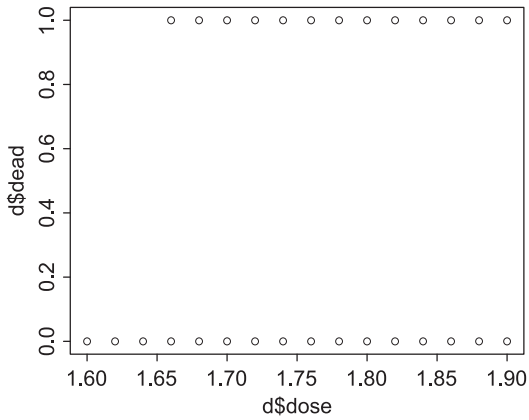

図 2.3　薬剤用量と生存・死亡の散布図

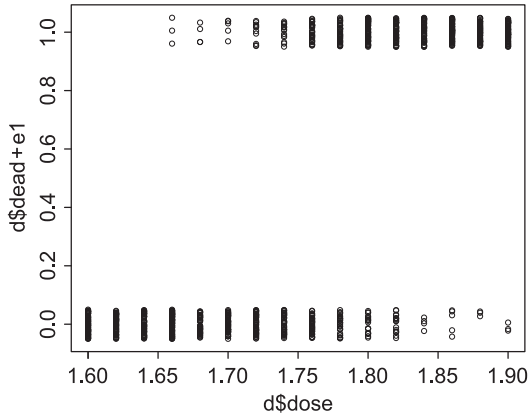

図 2.4　薬剤用量と生存・死亡の散布図：乱数で散らした場合

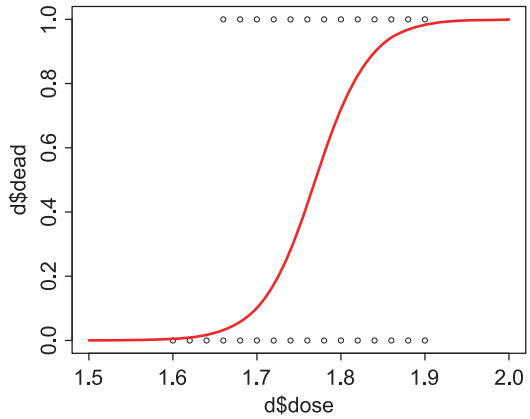

図 2.5　ロジスティック回帰による死亡確率の予測曲線

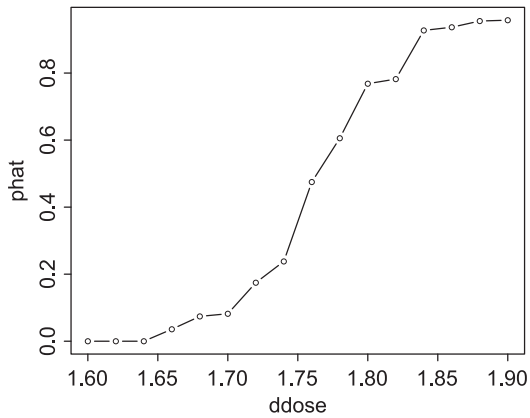

図 2.6　薬剤用量と死亡割合のグラフ

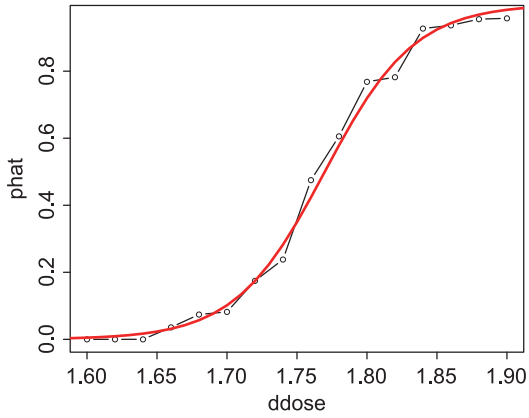

図 2.7　各薬剤用量に対する死亡割合のグラフとロジスティック回帰曲線

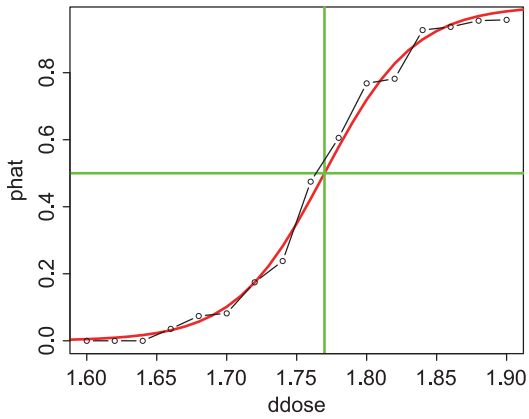

図 2.8　図 2.7 に `phat=0.5`, `ddose=1.77` の直線を追加したグラフ

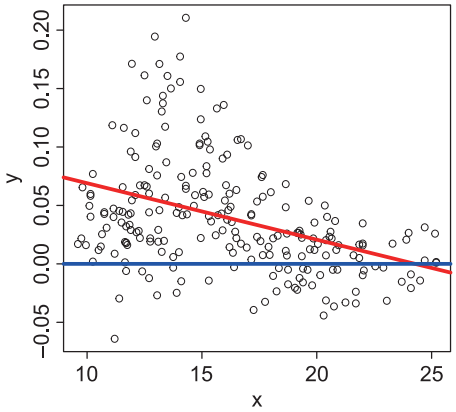

図 3.1　北アメリカの男性 261 名の年齢 (x) と骨密度の相対変化 (y)．回帰直線を赤色，相対変化なしを表す y=0 を青色で示す．

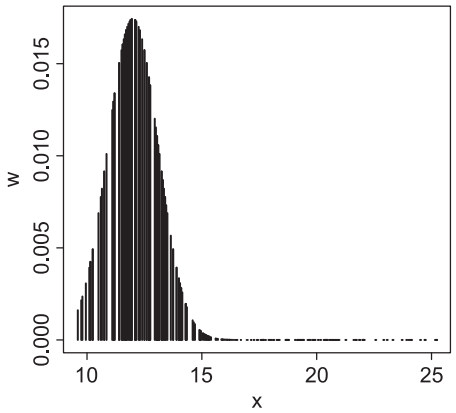

図 3.2　12 歳の周りの重み関数

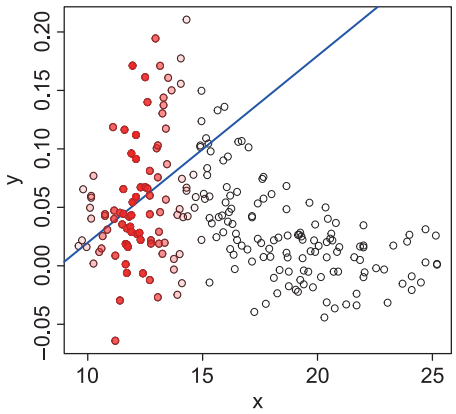

図 3.3　12 歳の周りの重み付き回帰．重みが大きい点ほど濃い赤色で示す．

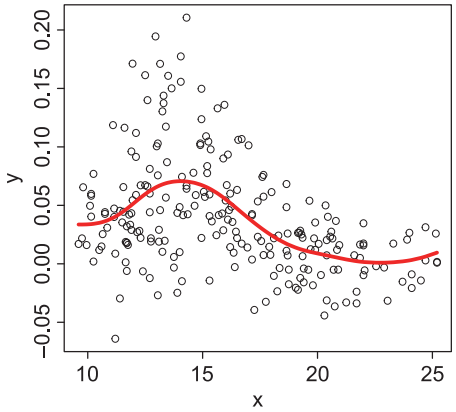

図 3.4　標準偏差を 1.1 とした場合の局所重み付き回帰による散布図平滑化．このとき残差平方和は 0.375.

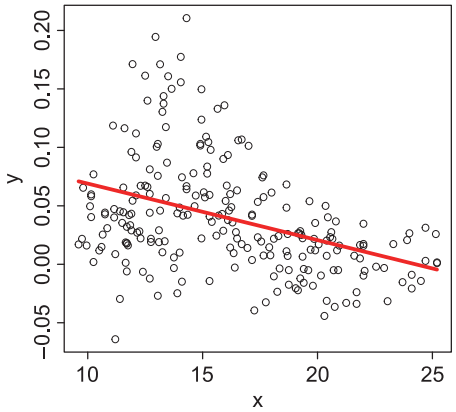

図 3.5　標準偏差を 100 とした重みによる局所重み付き回帰の適合曲線．このとき残差平方和は 0.453 と大きい.

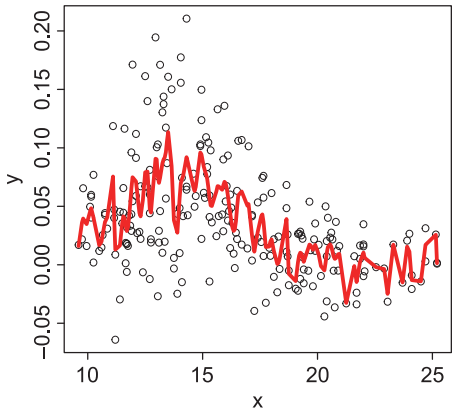

図 3.6　標準偏差を 0.1 とした重みによる局所重み付き回帰の適合曲線．このとき残差平方和は 0.265 と小さい.

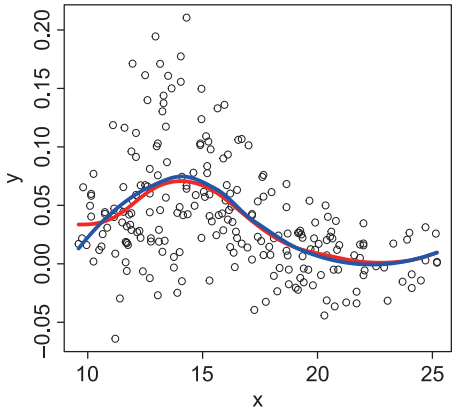

図 3.7　局所多項式回帰による散布図平滑化．`loess` 関数による適合曲線を青色で，標準偏差を 1.1 とした場合の局所重み付き回帰の適合曲線を赤色で示す．

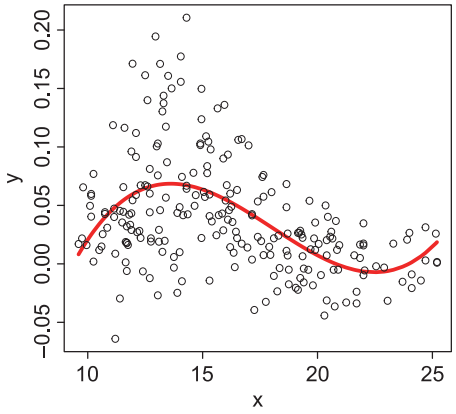

図 3.8　3 次曲線による適合曲線

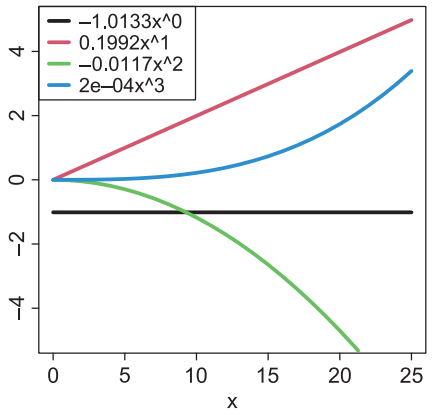

図 3.9　3 次曲線 $y = -1.013 + 0.199x - 0.012x^2 + 0.0002x^3$ における 4 つの基底関数と回帰係数の積の関数，すなわち，$y = -1.013$, $y = 0.199x$, $y = -0.012x^2$ および $y = 0.0002x^3$．凡例には，回帰係数と基底関数の積を示した．

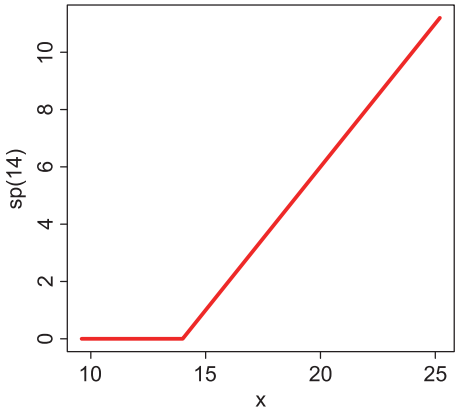

図 3.10　$x = 14$ を節点とする 1 次のスプライン関数

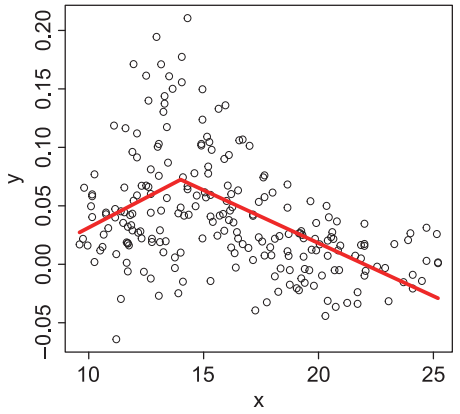

図 3.11　$x = 14$ を節点とする 1 次のスプライン基底関数を加えた適合曲線

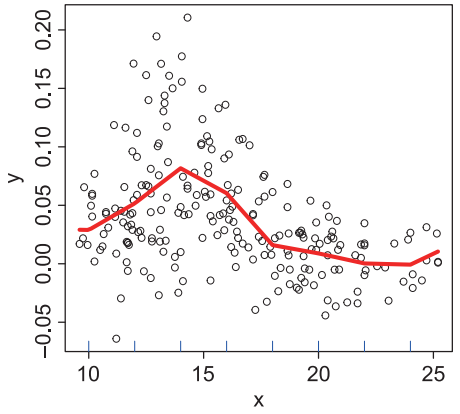

図 3.12　8 つの節点によるスプライン回帰．なお，説明変数は定数項と **x** を合わせて 10 個となる．

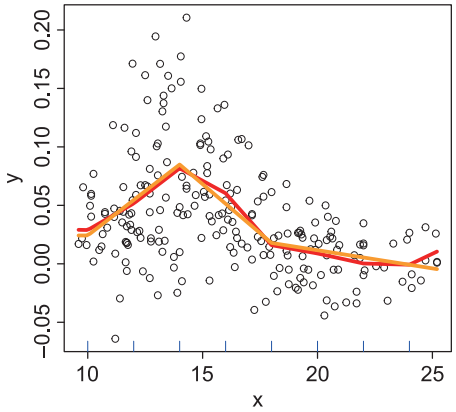

図 3.13　AIC を用いた変数選択の結果によるスプライン曲線．変数選択前の 10 個の説明変数を用いた曲線を赤色，最適な 4 個の説明変数による曲線をオレンジ色で示す．

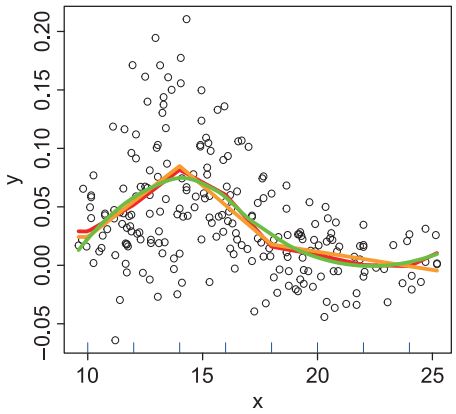

図 3.14　`loess` 関数による曲線を緑色で図 3.13 に追加

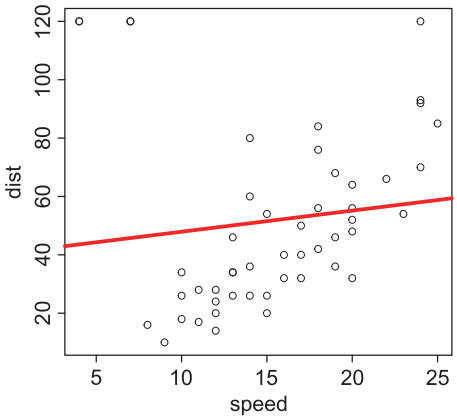

図 3.15　4 点の外れ値がある場合にあてはめた通常の回帰直線

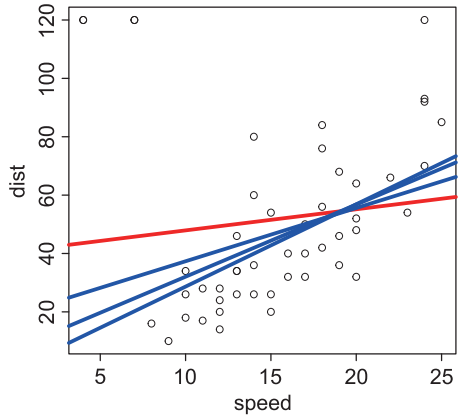

図 3.16　4 点の外れ値がある場合に残差を利用した重み付き回帰を 3 回適用した場合．通常の回帰直線を赤色で，3 度の重み付き回帰直線を青色で示す．

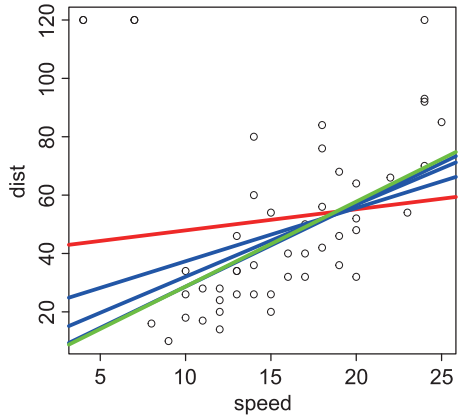

図 3.17　ロバスト回帰を適用した場合．通常の回帰直線を赤色，3 度の重み付き回帰を青色，`rlm` 関数によるロバスト回帰を緑色で示す．

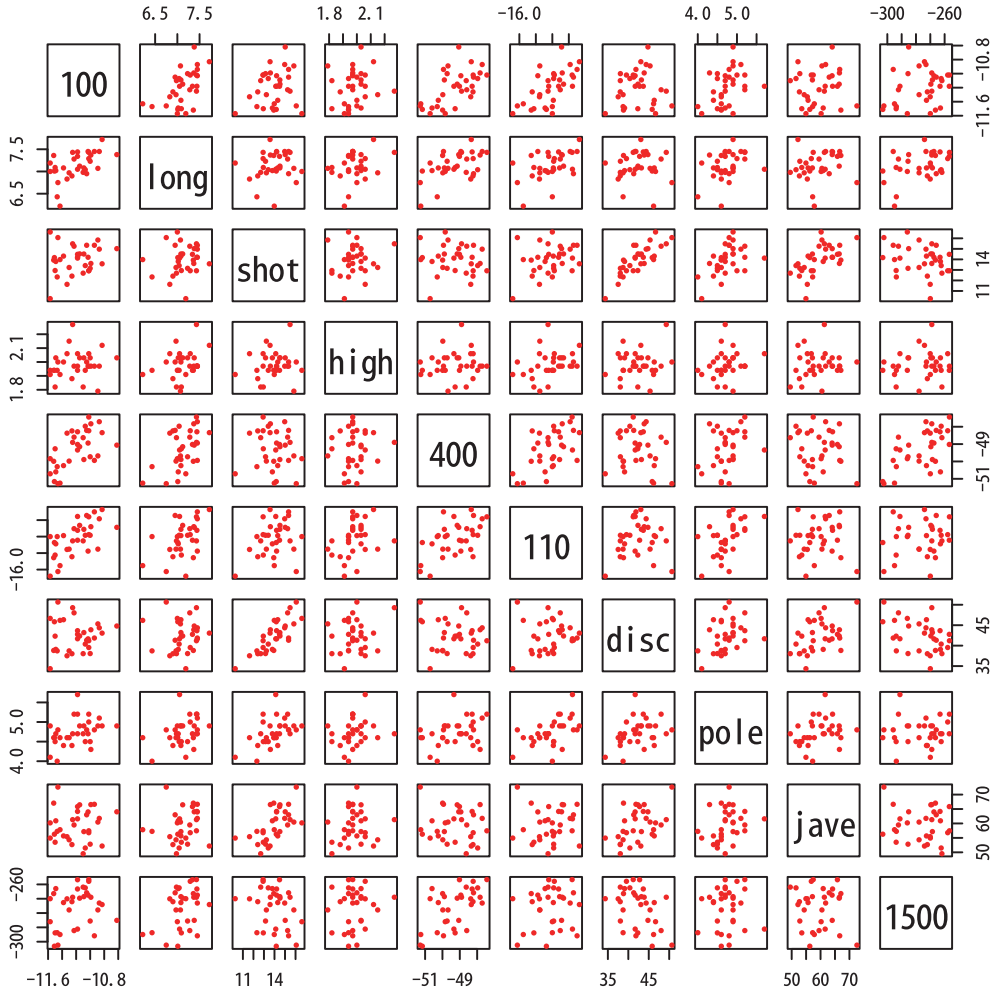

図 4.1　十種競技の散布図行列

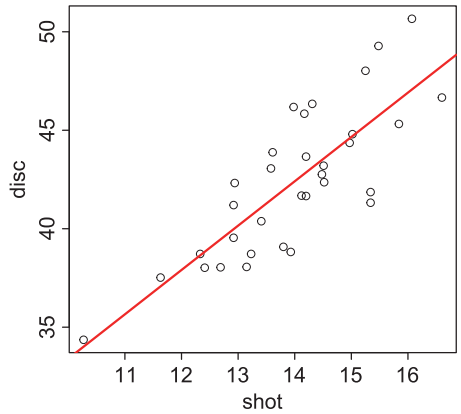

図 4.2　横軸を `shot` (砲丸投) 縦軸を `disc` (円盤投) とした散布図と回帰直線

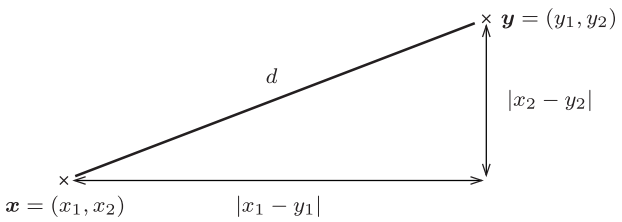

図 4.3　2 次元座標におけるピタゴラスの定理の模式図

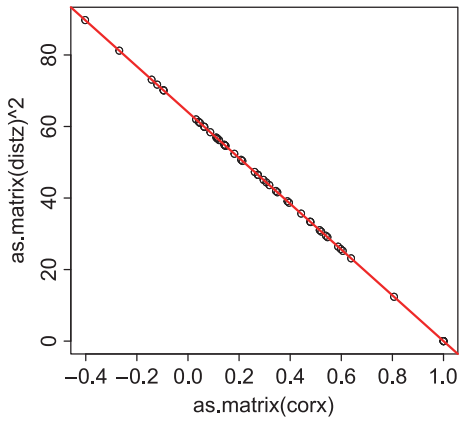

図 4.4　標準化データにおける相関係数とユークリッド距離の 2 乗

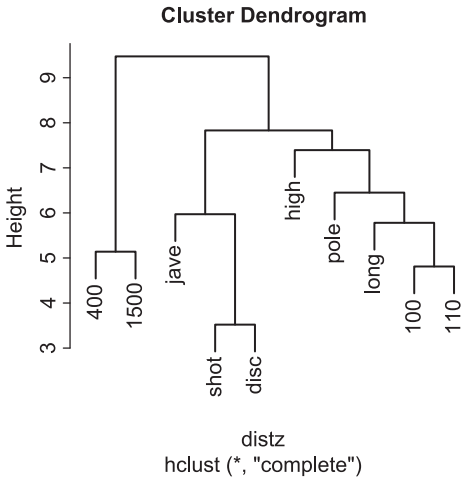

図 4.5　階層型クラスター分析によるデンドログラム

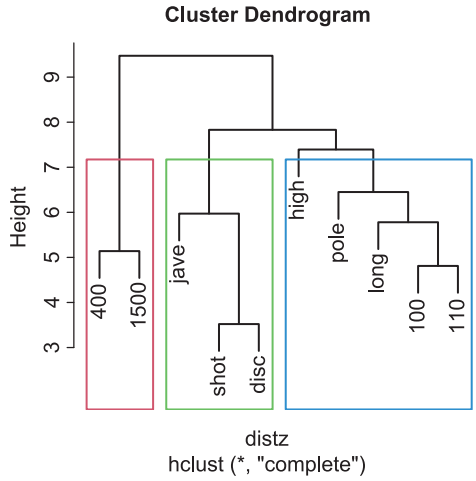

図 4.6　階層型クラスター分析による変数の分類

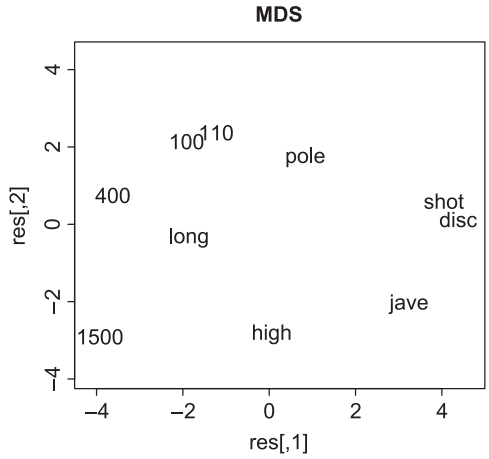

図 4.7　距離行列から復元された多次元尺度による十種競技の配置

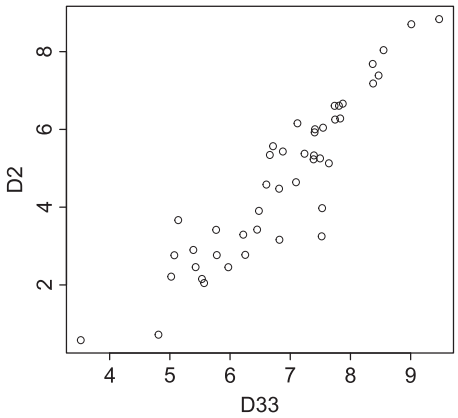

図 4.8　十種競技の 33 次元空間上の種目間距離と多次元尺度による 2 次元空間の種目間距離の対応

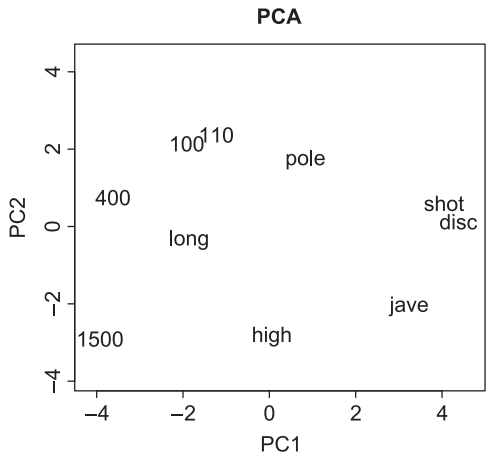

図 4.9　主成分分析による十種競技の 2 次元配置．多次元尺度法の結果と一致する．

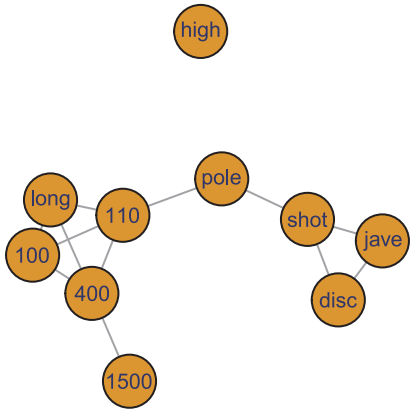

図 4.10　十種競技の無向グラフ

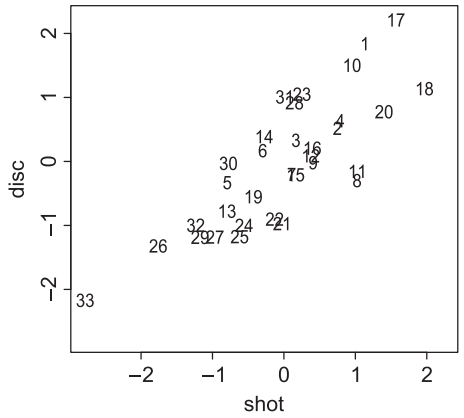

図 4.11　標準化された `shot` と `disc` を用いた 33 選手の散布図

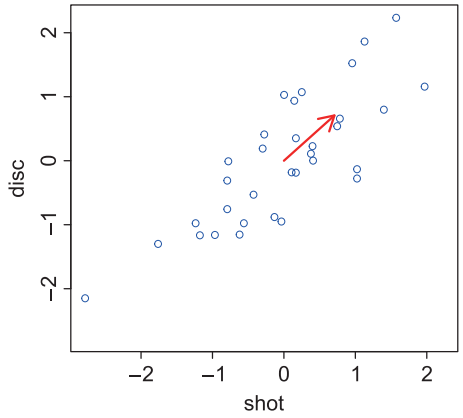

図 4.12　2 次元散布図を射影する軸の向き

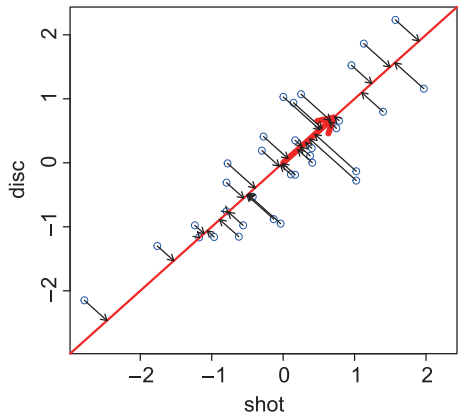

図 4.13　各点から矢印が示す軸に下した垂線の足

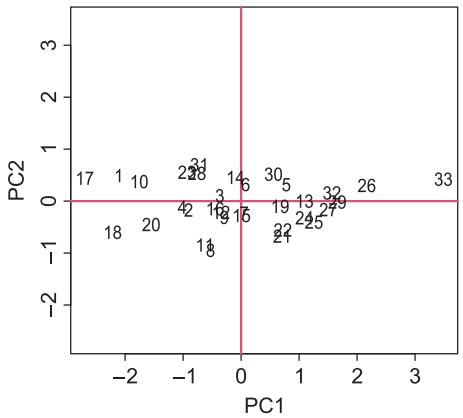

図 4.14　2 次元データ (`shot`, `disc`) に対する主成分分析

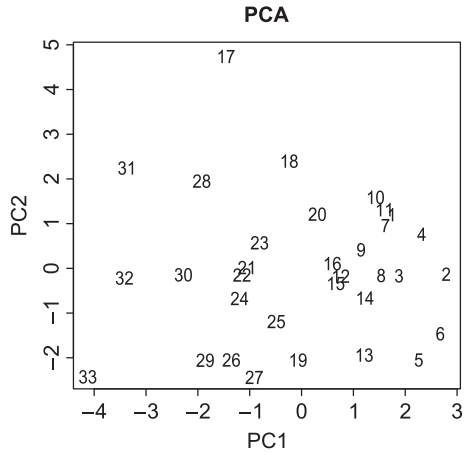

図 4.15　10 次元データに対する主成分分析の結果

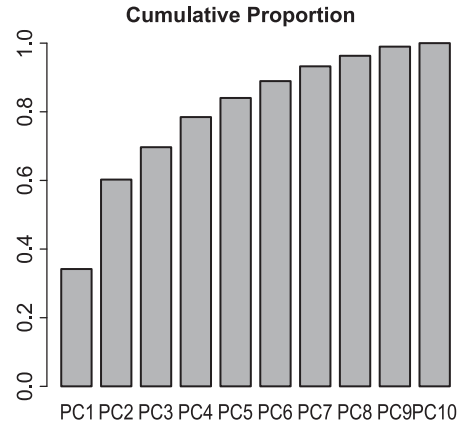

図 4.16　10 次元データに対する主成分分析の累積寄与率

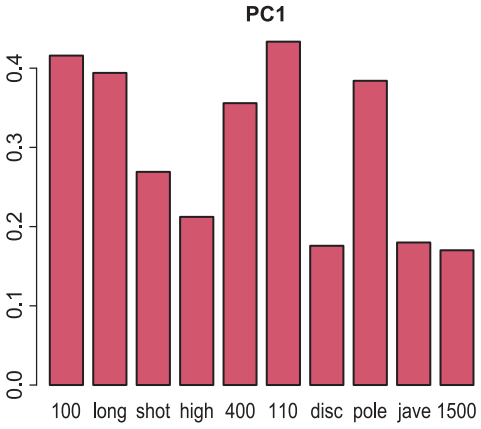

図 4.17　第 1 主成分の係数

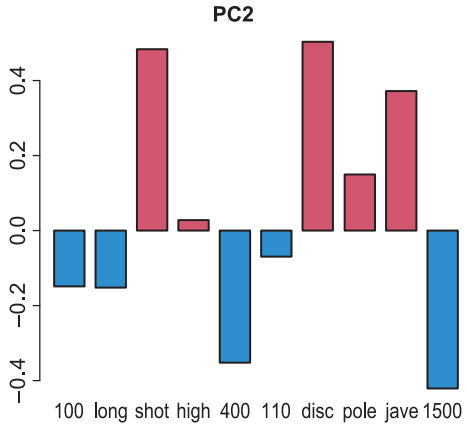

図 4.18　第 2 主成分の係数

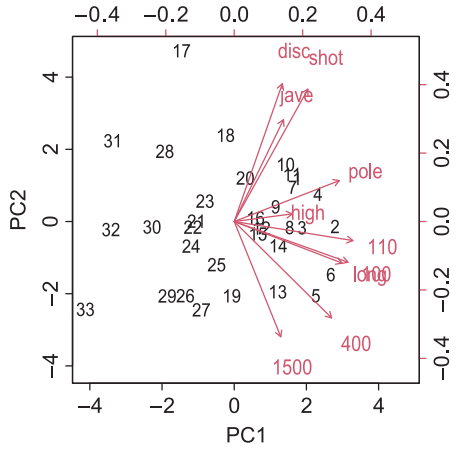

図 4.19　バイプロット

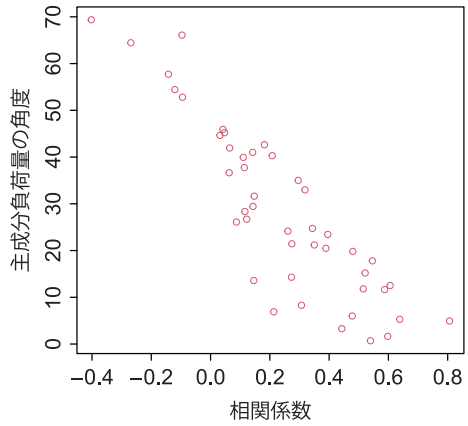

図 4.20　相関係数と図 4.19 における係数ベクトルの矢印どうしがなす角度

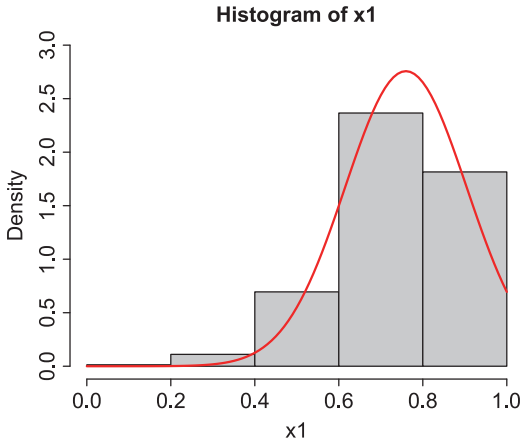

図 5.1　死亡グループの予測確率 $\hat{p}_i$ のヒストグラム (密度表示) と正規分布の重ね合わせ

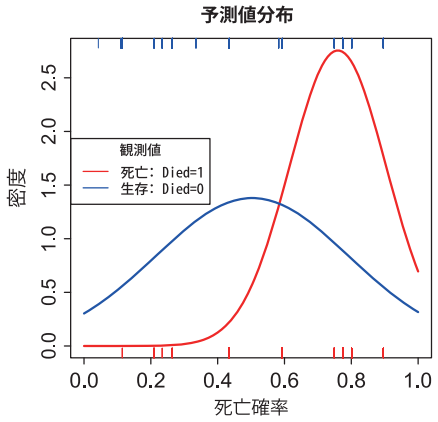

図 5.2　予測確率 $\hat{p}_i$ の分布に対する正規分布あてはめ (赤: 死亡グループ, 青: 生存グループ)

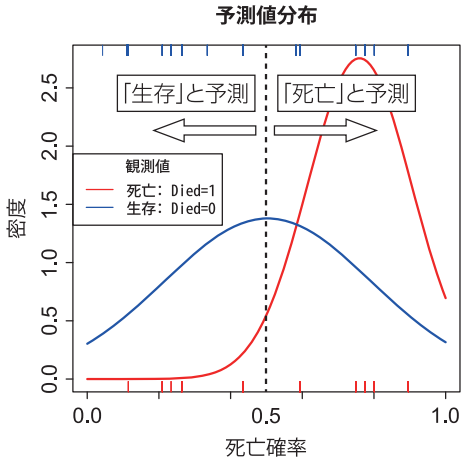

図 5.3　予測確率の分布の正規分布あてはめ (図 5.2 と同じ) と閾値 0.5 の判別ルール

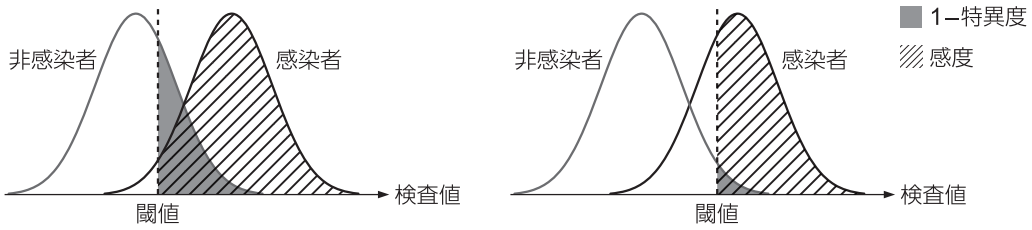

図 5.4　感染者と非感染者の検査値の分布と閾値の設定

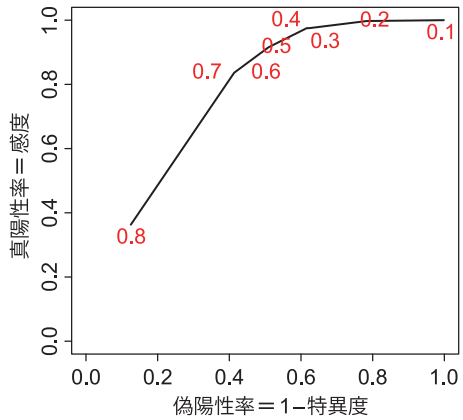

図 5.5　ロジスティック判別：閾値を変化させたときの感度と 1 − 特異度 のグラフ

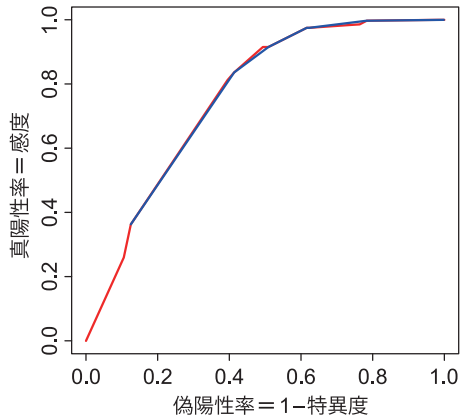

図 5.6　ロジスティック判別：ROC 曲線と図 5.5 の重ね描き

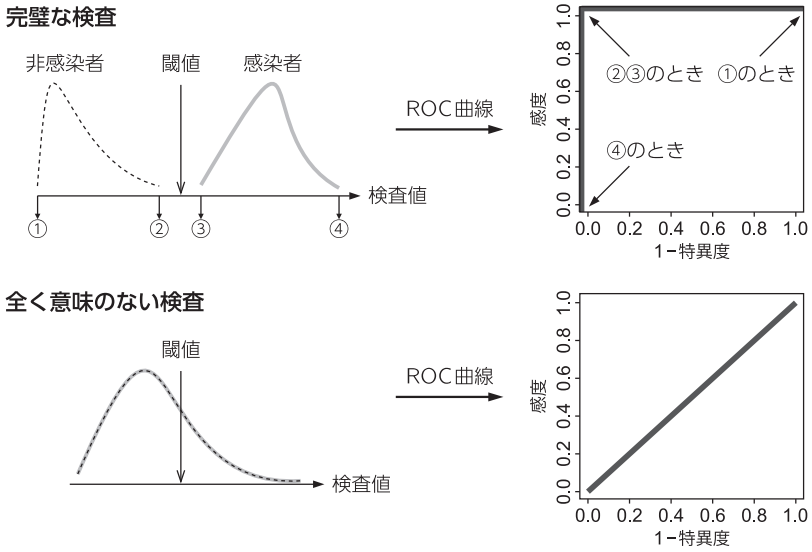

図 5.7　完璧な検査とまったく意味のない検査の分布と対応する ROC 曲線

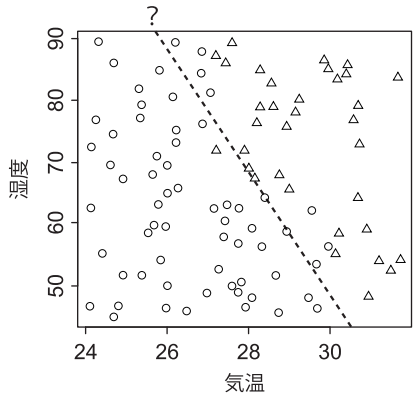

図 5.8　気温と湿度の組み合わせによって暑いと感じた (三角) か否か (丸)

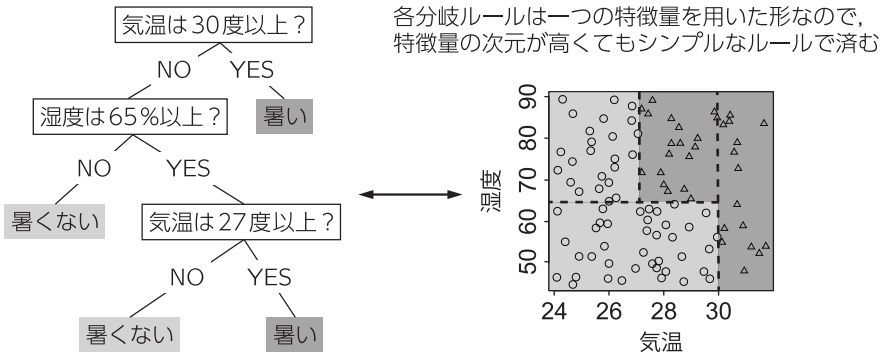

図 5.9　気温と湿度による決定木の作成例

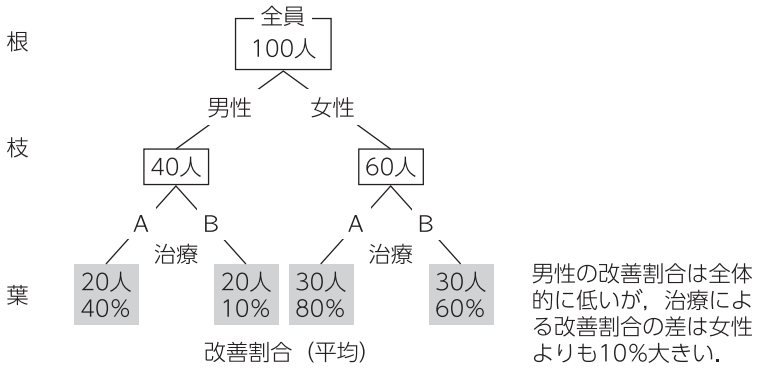

図 5.10　改善の有無を目的とした決定木の作成例

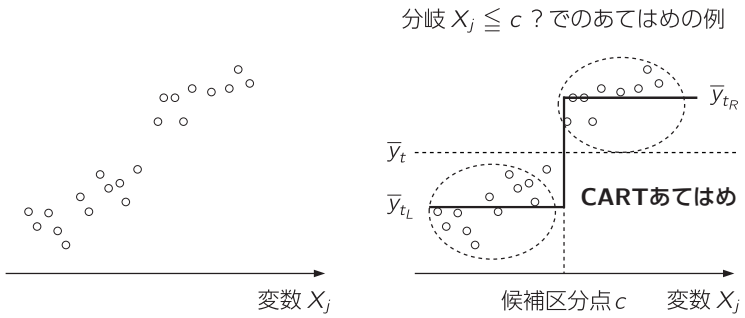

図 5.11　決定木 (CART) の分岐ルールの区分点の選び方の例

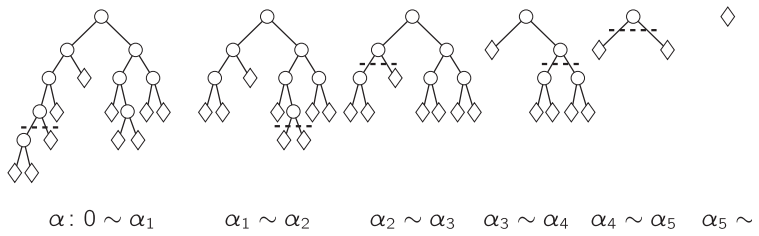

図 5.12　いくつかの決定木の例：木の大きさの選び方

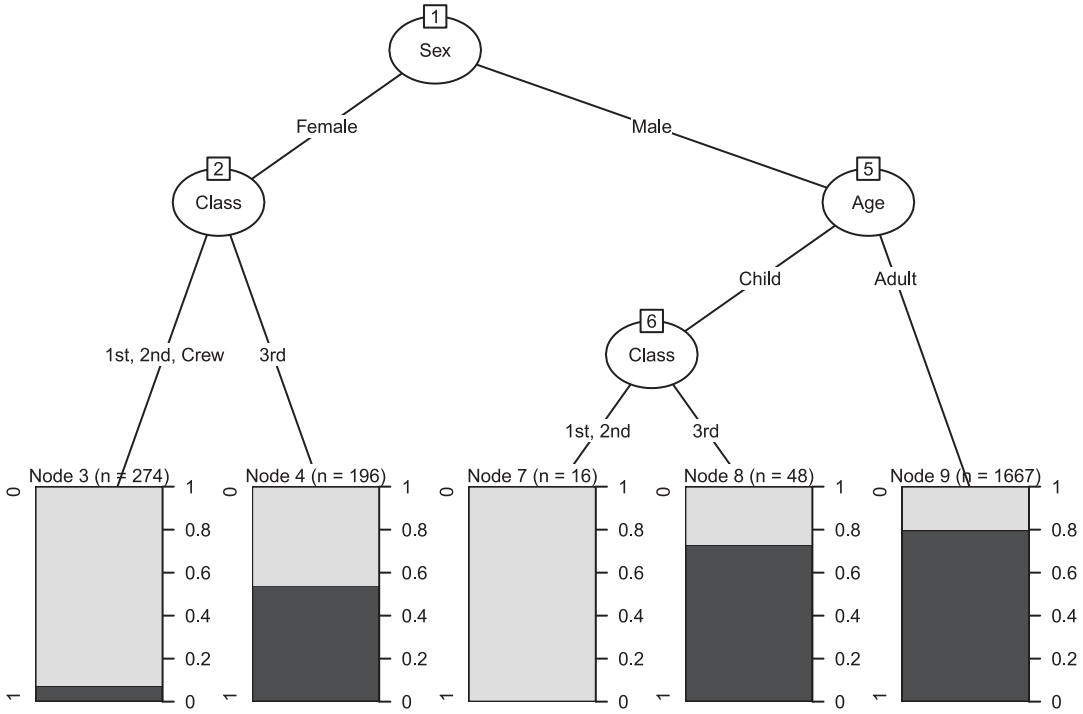

図 5.13　タイタニック号データに対する分類木の適用例

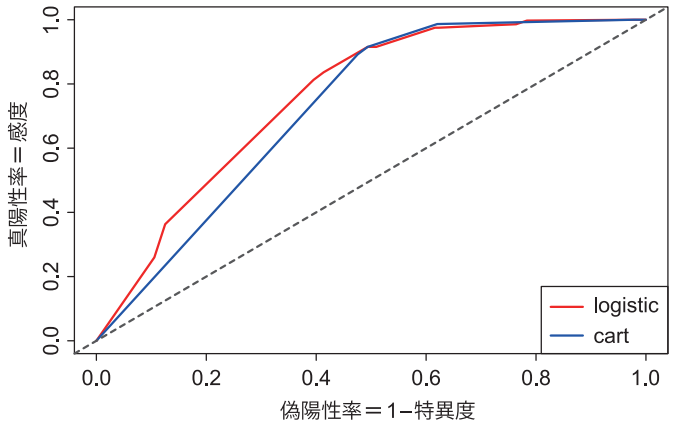

図 5.14　ロジスティック判別による ROC 曲線 (赤) と決定木による ROC 曲線 (青)

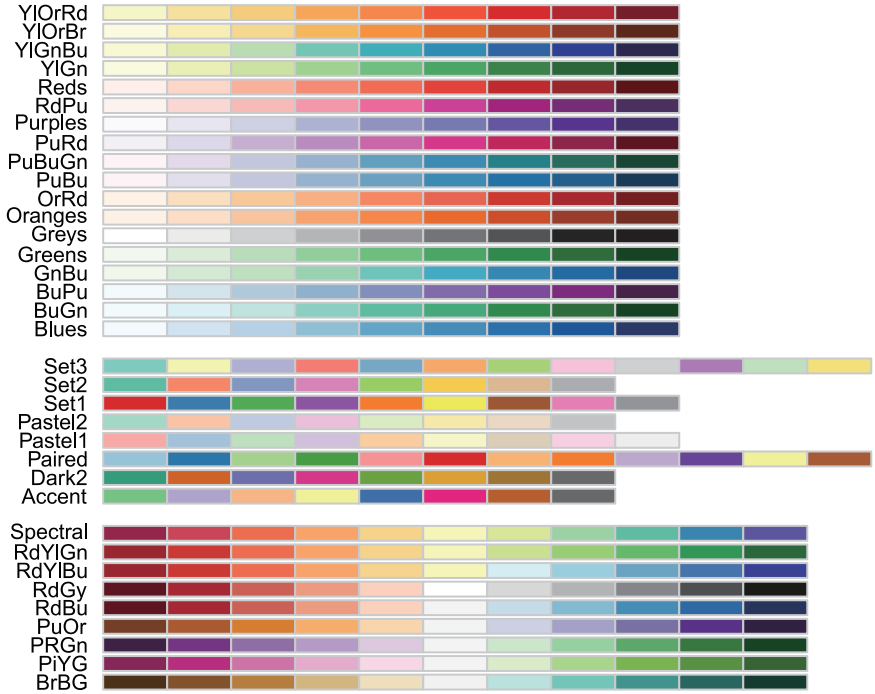

図 6.1　`RColorBrewer` パッケージに収録された色見本．`display.brewer.all()` を実行．

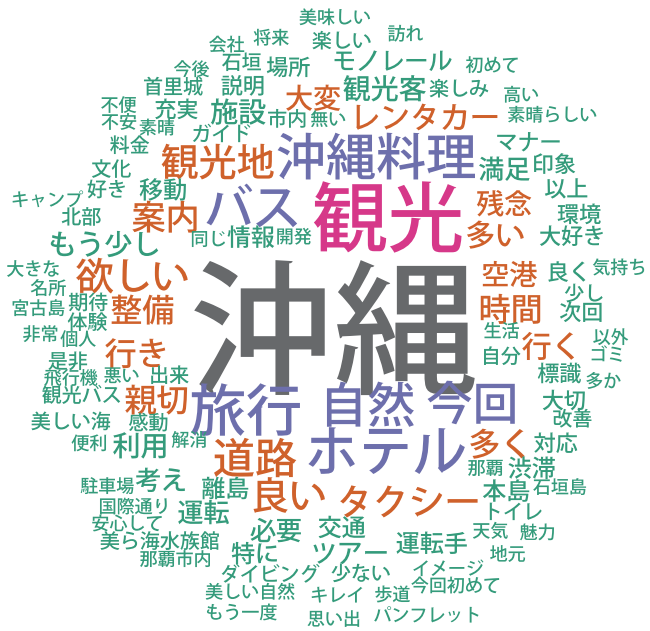

図 6.2　頻出単語によるワードクラウド

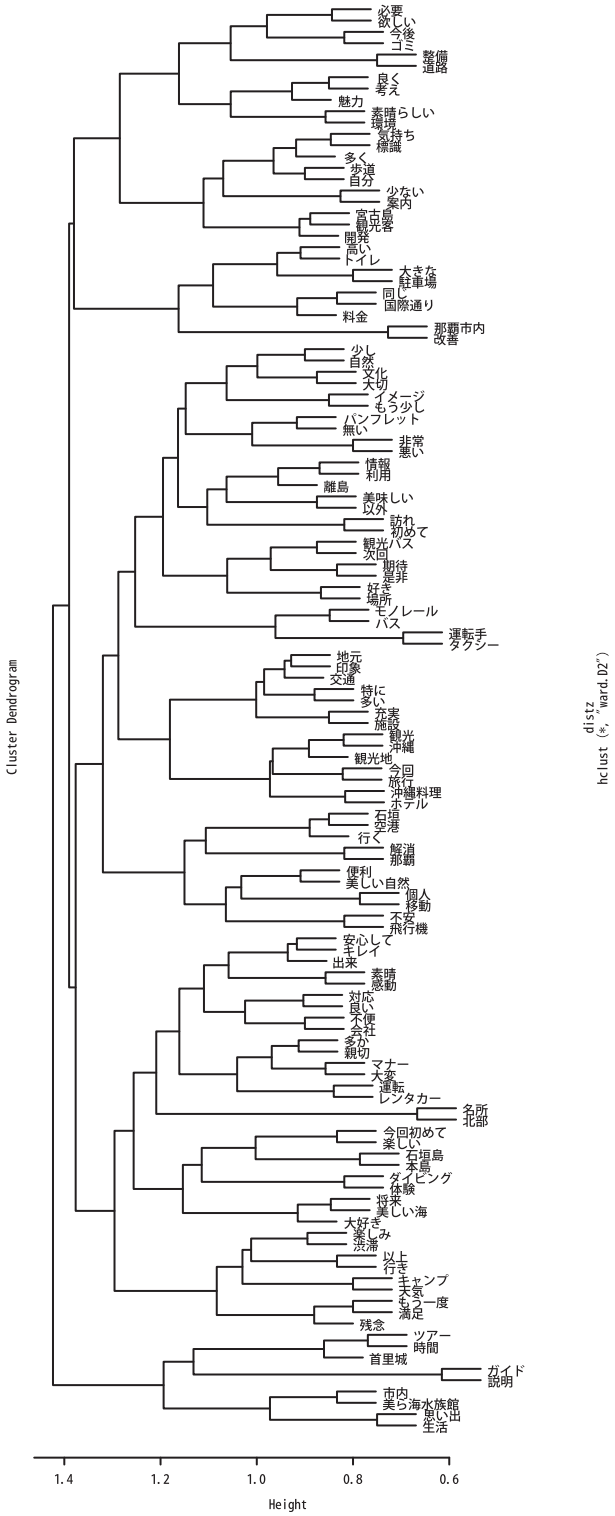

図 6.3　階層型クラスター分析

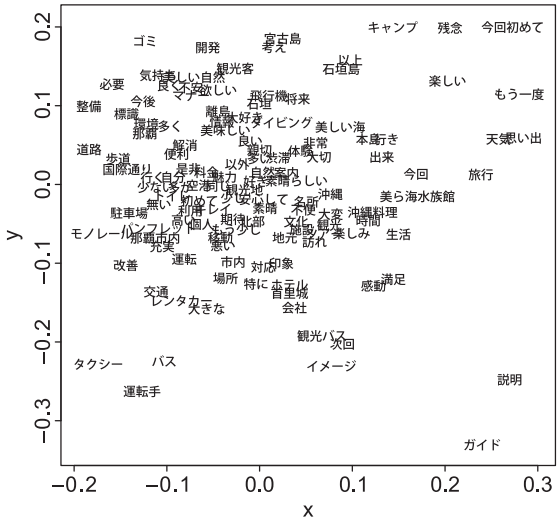

図 6.4　多次元尺度法による視覚化

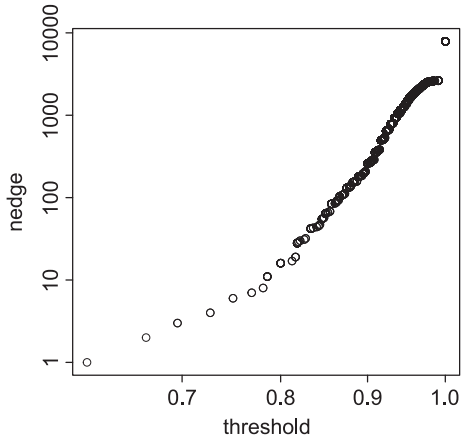

図 6.5　ジャッカード距離の閾値とエッジ数の両対数図

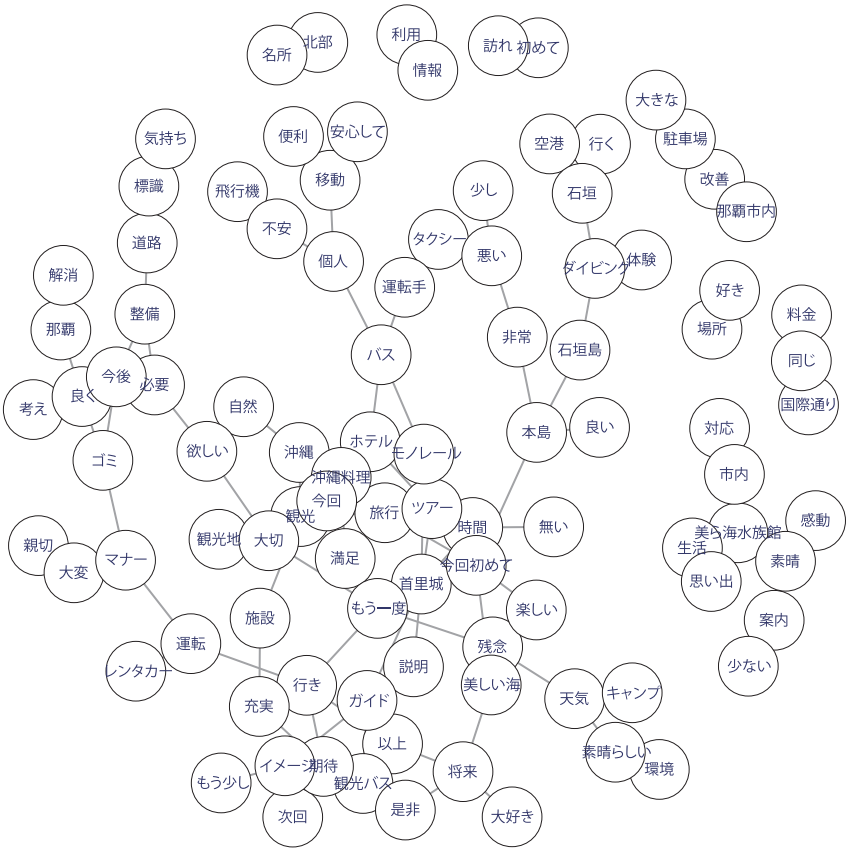

図 6.6　隣接行列を基に作成されたネットワーク図

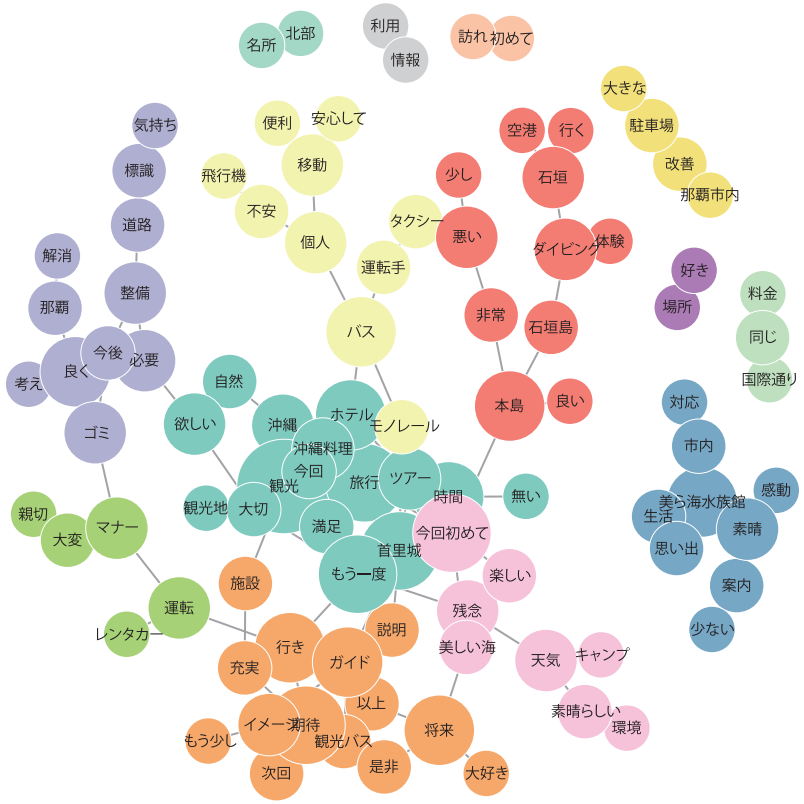

図 6.7　共起ネットワーク

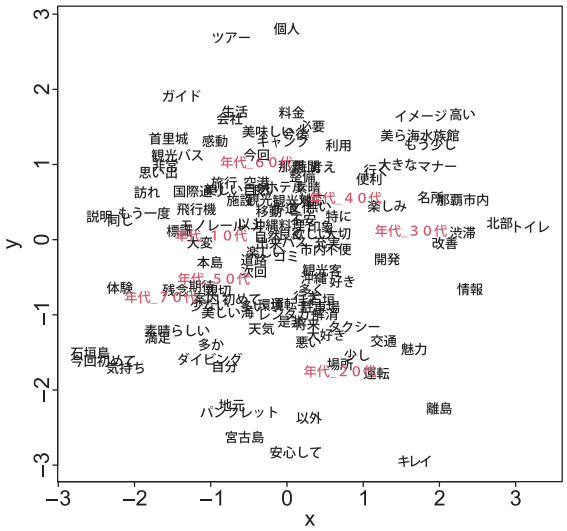

図 6.8　年代と頻出単語のクロス集計表に対する対応分析

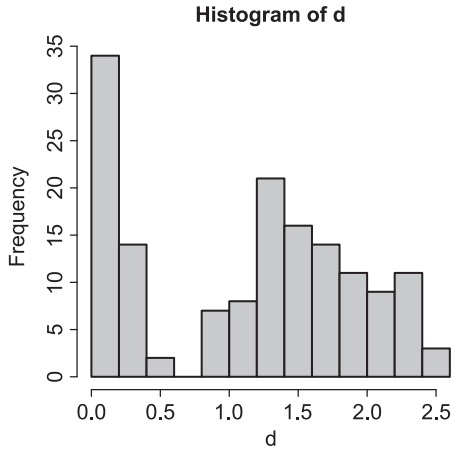

図 7.1　`iris` データの `Petal.Width` のヒストグラム

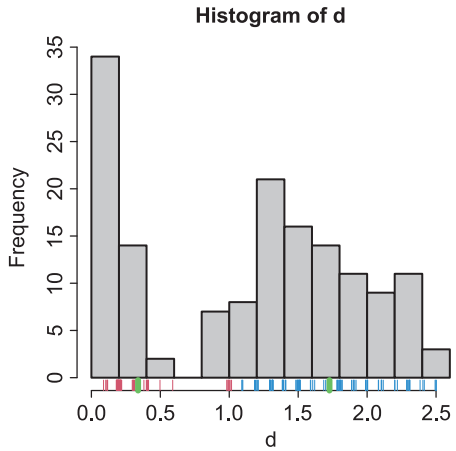

図 7.2　`Petal.Width` の *k*-means 法による分類結果

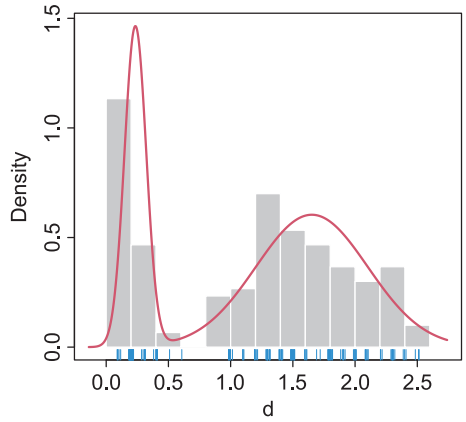

図 7.3　`Petal.Width` に対して混合正規分布をあてはめた結果

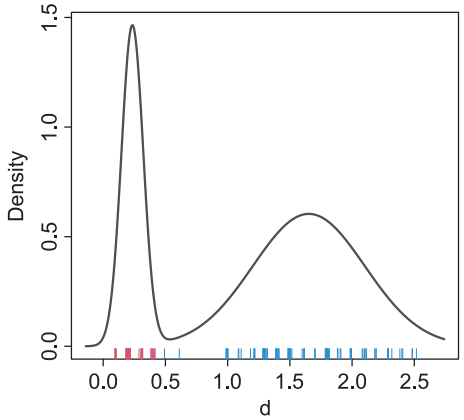

図 7.4　`Petal.Width` の混合正規分布による分類結果

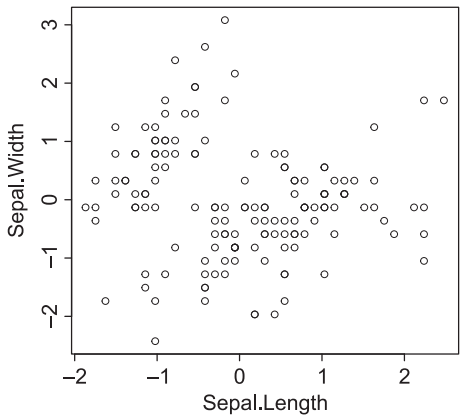

図 7.5　`iris` データの (`Sepal.Length`, `Sepal.Width`) による散布図

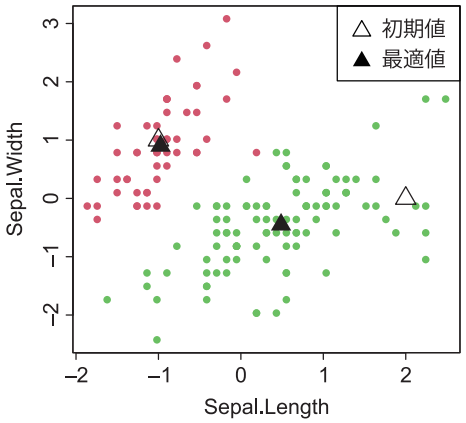

図 7.6　(`Sepal.Length`, `Sepal.Width`) の *k*-means 法による分類結果 (初期値その 1)

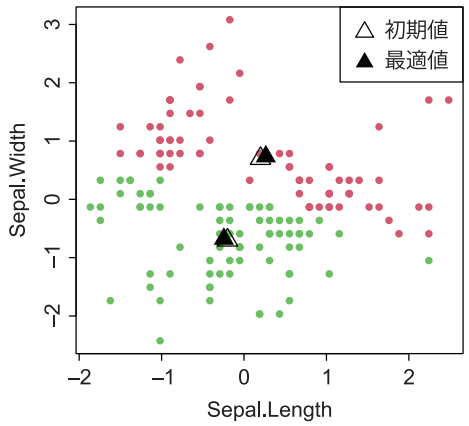

図 7.7 (`Sepal.Length`, `Sepal.Width`) の *k*-means 法による分類結果 (初期値その 2)

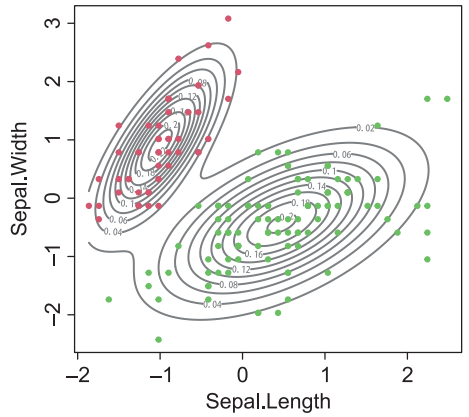

図 7.8 (`Sepal.Length`, `Sepal.Width`) の *k*-means 法による分類結果

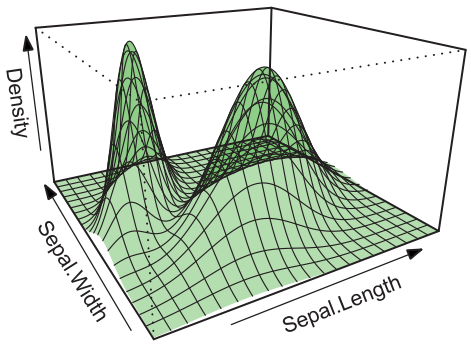

図 7.9 (`Sepal.Length`, `Sepal.Width`) にあてはまった混合正規分布の密度関数の鳥観図

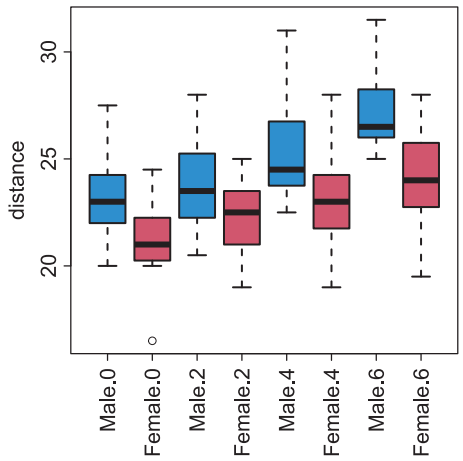

図 7.10　`Orthodont` データにおける `distance` の年齢別性別の箱ひげ図

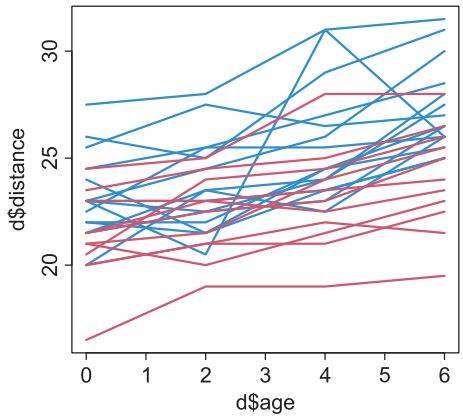

図 7.11　`Orthodont` データにおける `distance` の個体別経時変化．男性は青色，女性を赤色で示す．

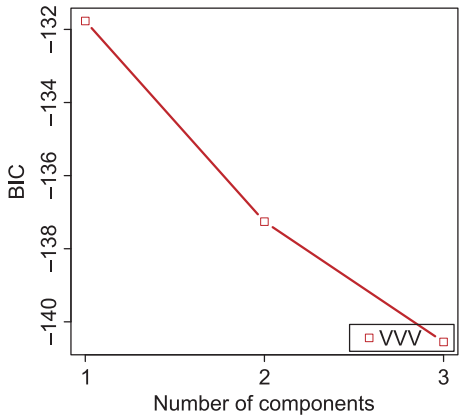

図 7.12　`distance` の経時変化に直線をあてはめて得られる切片と傾きの 2 次元データ `bmat` に対して混合正規分布をあてはめたときのコンポーネント数に対するモデル選択基準 `BIC`．`BIC` が最も小さいコンポーネント数 3 が最良のモデルとなる．

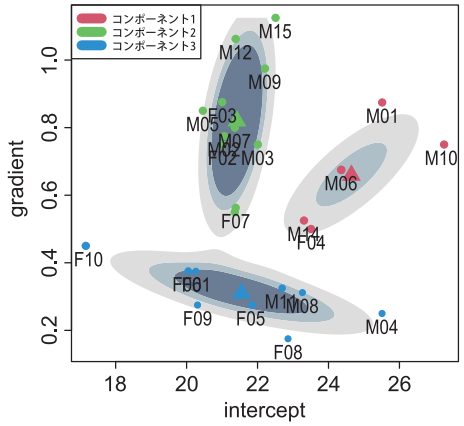

図 7.13　切片と傾きの 2 次元データ `bmat` に対してコンポーネント数 3 の混合正規分布をあてはめた結果．コンポーネントごとの平均を塗りつぶしの三角形 (`pch=17`) で示す．

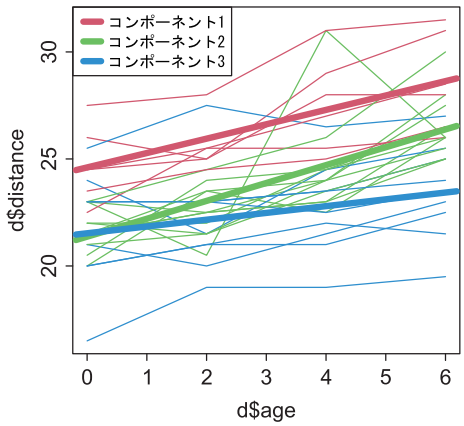

図 7.14　混合正規分布による折れ線の分類結果．コンポーネントごとの平均を太線で示す．

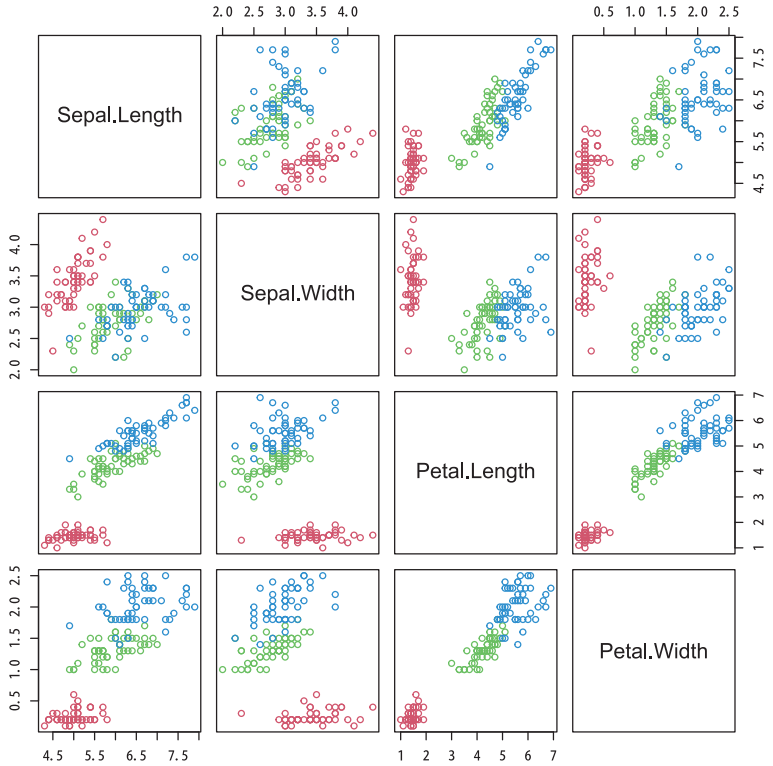

図 7.15　`iris` データの散布図行列．`Species` の値で色分けをした．

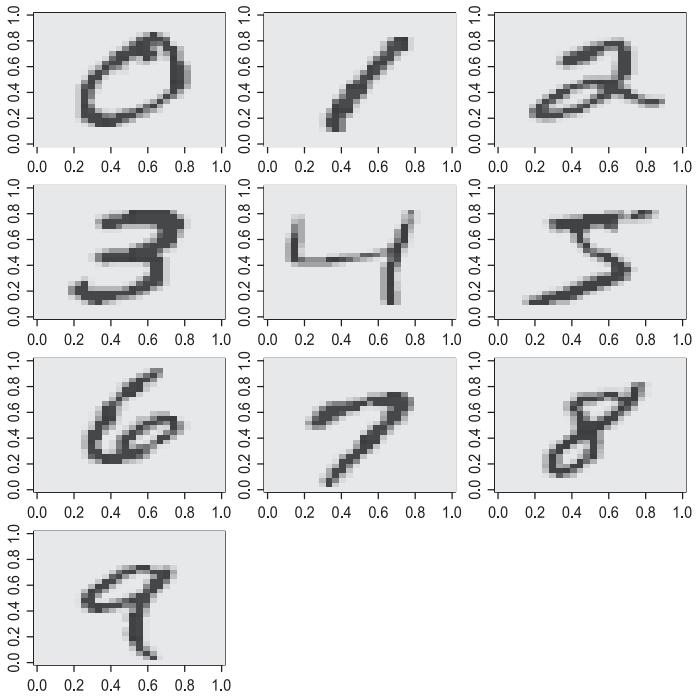

図 7.16　手書き数字の画像データ．画像は縦 28 個，横 28 個の四角に該当するピクセルという単位で構成されている．CSV データの 1 行には正解ラベルとして 0 から 9 の数字に続いて，各ピクセルの色を示す 0 から 255 までの数値が 28 × 28 = 784 ピクセル分入力されている．ここで，色はグレースケールで，0 が白，255 が黒を示す．

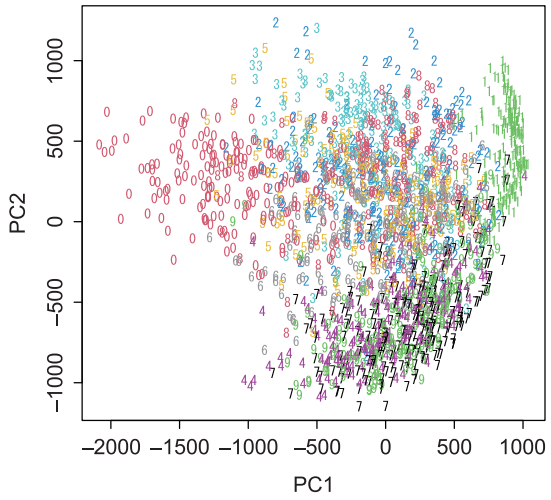

図 7.17　手書き数字の画像データに対して主成分分析を適用した結果．正解ラベルの数字をプロットし，数字ごとに色分けした．

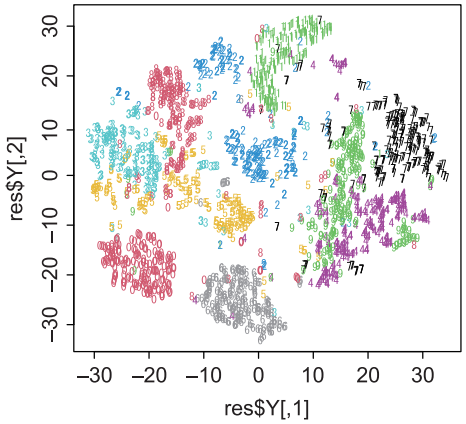

図 7.18　手書き数字の画像データに対して t-SNE 法を適用した結果．正解ラベルの数字をプロットし，数字ごとに色分けした．

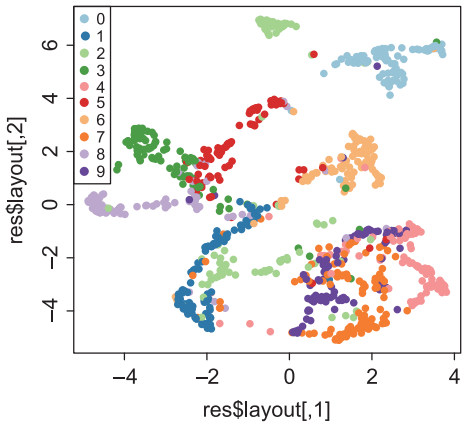

図 7.19　手書き数字の画像データに対して UMAP 法を適用した結果．正解ラベルの数字ごとに色分けした．

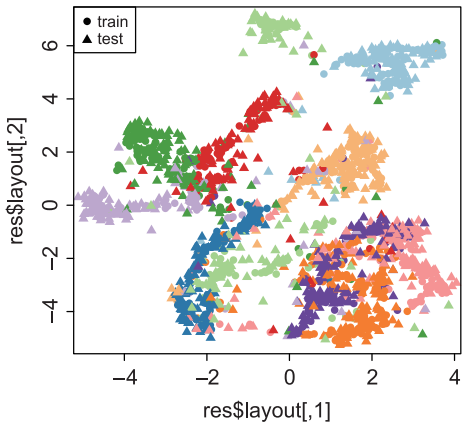

図 7.20　手書き数字の画像データに対して UMAP 法を訓練データに適用し，検証データに対して予測を行った結果の配置．

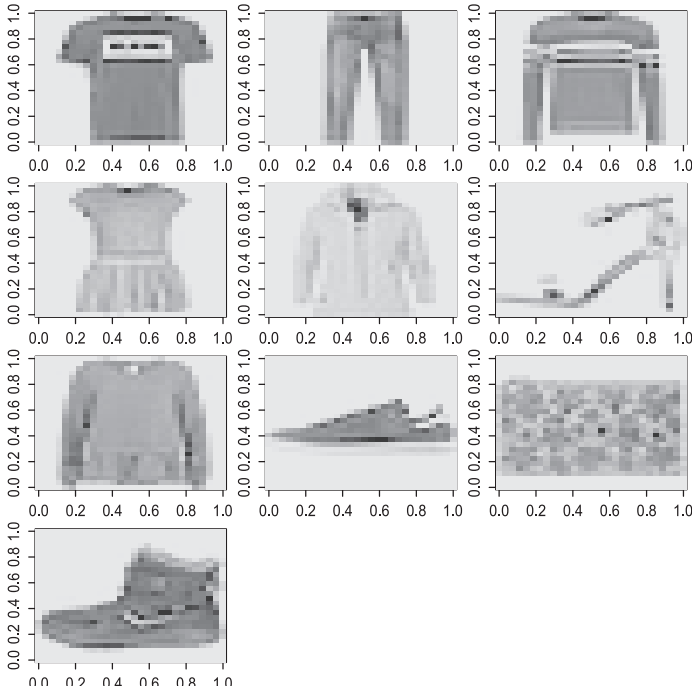

図 7.21　ファッションアイテムの画像データ

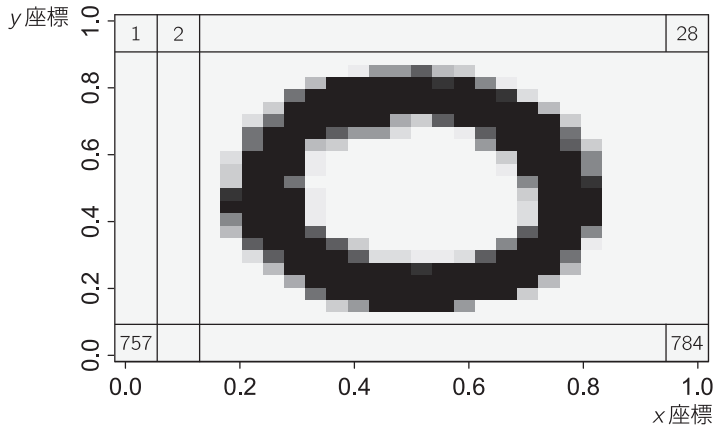

図 8.1　handwrite0.csv データの `image` 関数の適用とピクセル構成のイメージ

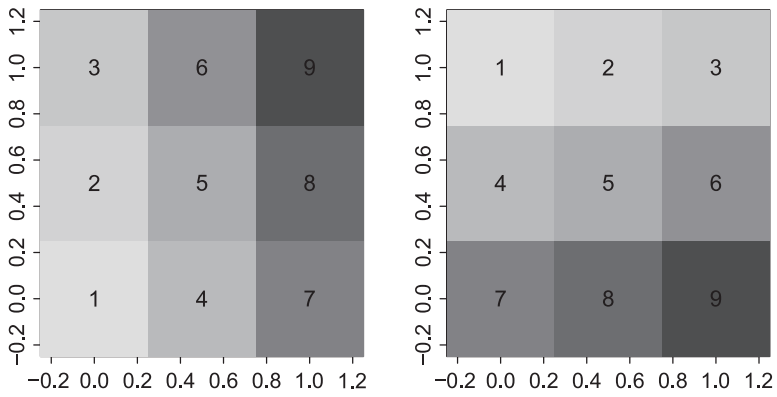

図 8.2　行列 `matrix(1:9,ncol=3,byrow=T)` に対する `image` 関数の 2 つの適用例：左図は `image(x)` の適用結果，右図は `image(t(x)[,3:1])` の適用結果.

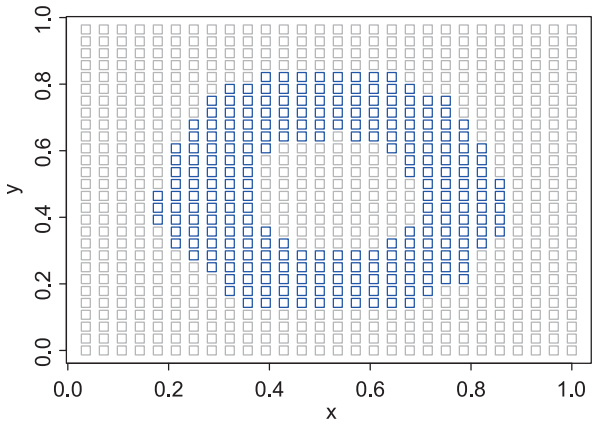

図 8.3　handwrite0.csv データ (変数 `z` の 2 値化) の可視化の例

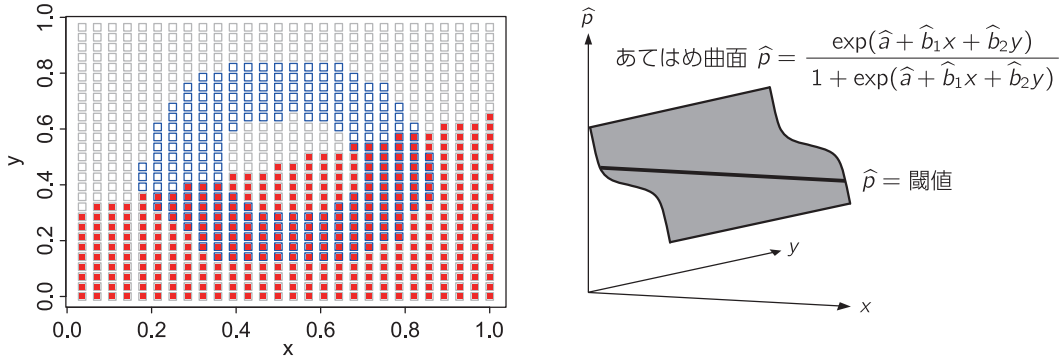

図 8.4　サンプルデータに対するロジスティック回帰 (8.1) による判別結果 (赤色)

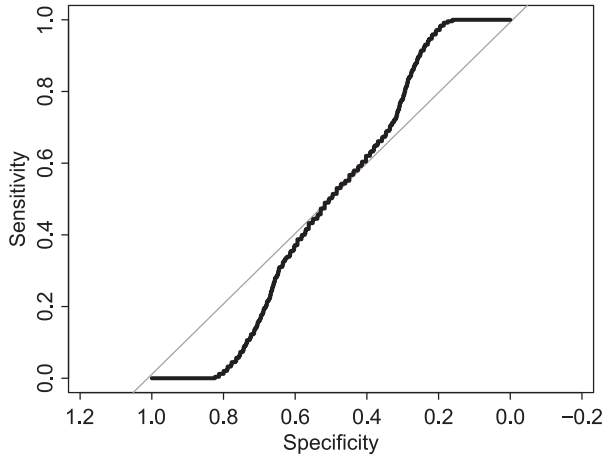

図 8.5　サンプルデータへのロジスティック回帰のあてはめの ROC 曲線

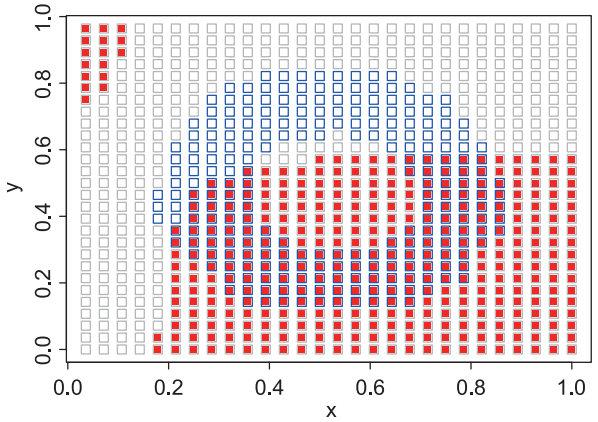

図 8.6　サンプルデータに対するロジスティック回帰 (8.6) による判別結果 (赤色)

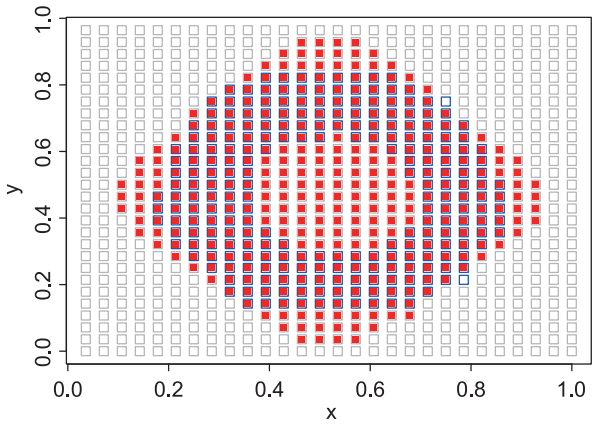

図 8.7　サンプルデータに対するロジスティック回帰 (8.5) による判別結果 (赤色)

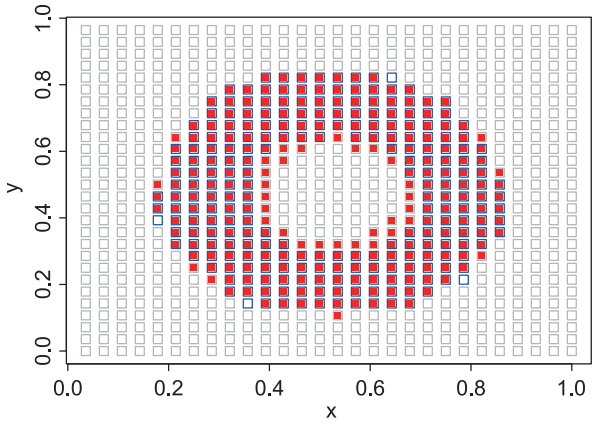

図 8.8　サンプルデータに対する SVM を用いた判別結果

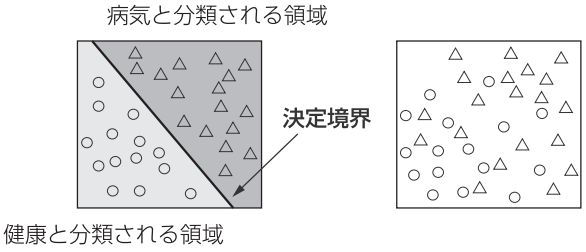

図 8.9　2 クラス分類：線形分離可能な状況 (左) とそうでない場合 (右)

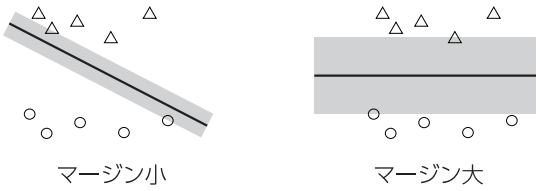

図 8.10　決定境界とマージン

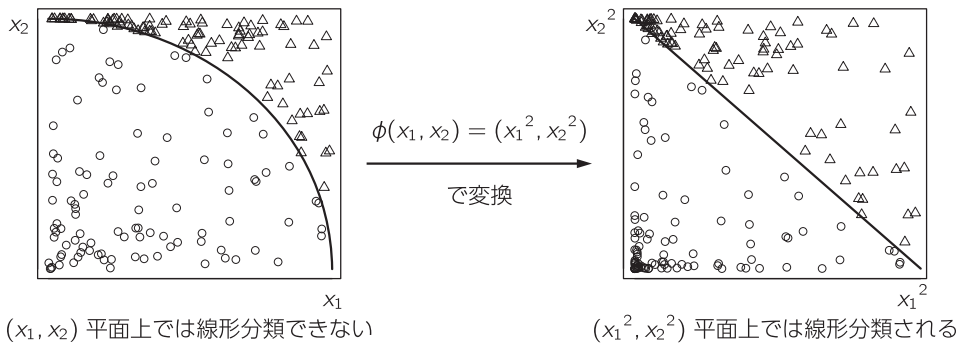

図 8.11　カーネル法による非線形分類の例示

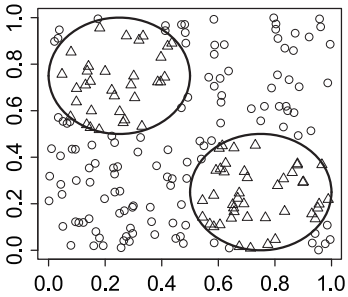

図 8.12　より複雑な決定境界が求められる場合の例示

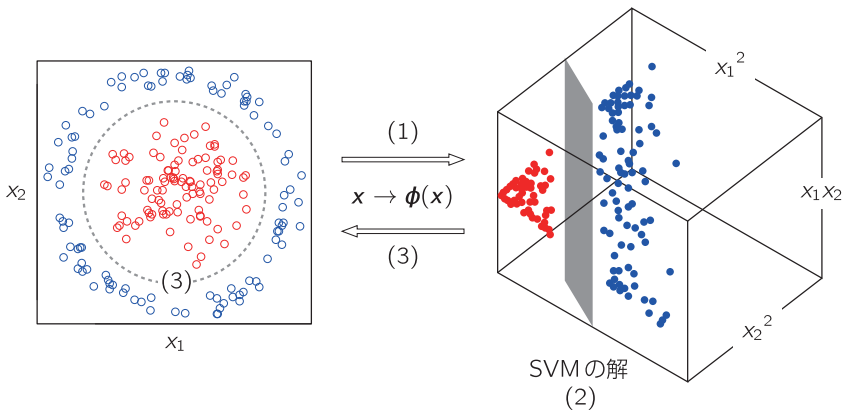

図 8.13　データの高次元の変換によって線形分離可能にできる仕組みの例示

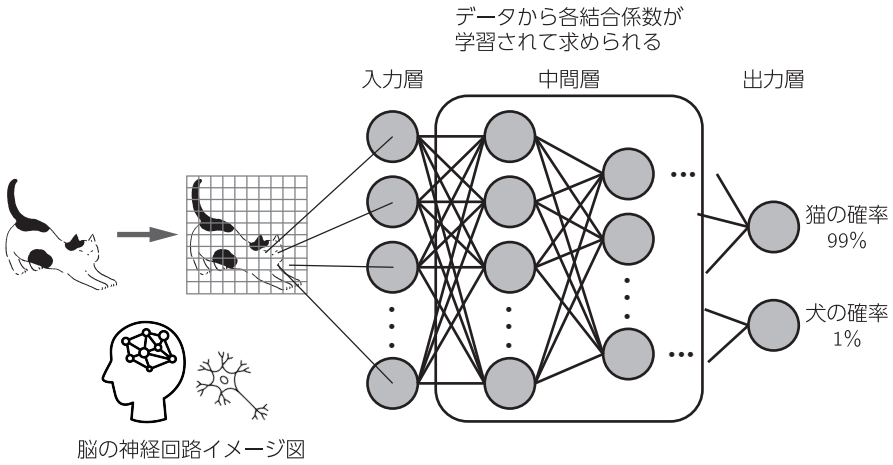

図 8.14　深層ニューラルネットのイメージ図

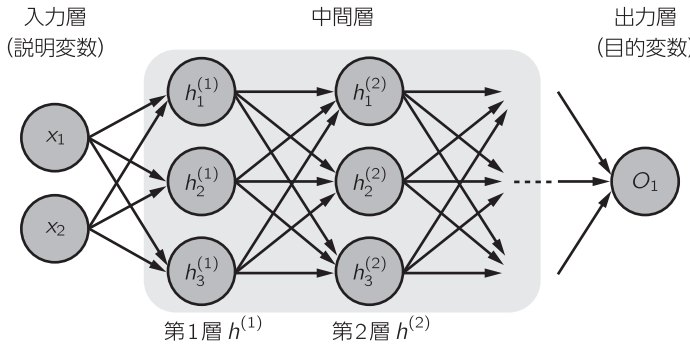

図 8.15　ニューラルネットワークによる予測モデルの模式図

第 1 層のユニット i の値(出力)を決める式

$$h_i^{(1)} = \frac{\exp\left(b_{i0}^{(1)} + b_{i1}^{(1)}x_1 + \cdots + b_{ip}^{(1)}x_p\right)}{1 + \exp\left(b_{i0}^{(1)} + b_{i1}^{(1)}x_1 + \cdots + b_{ip}^{(1)}x_p\right)} = f\left(b_{i0}^{(1)} + \sum_{j=1}^{p} b_{ij}^{(1)}x_j\right)$$

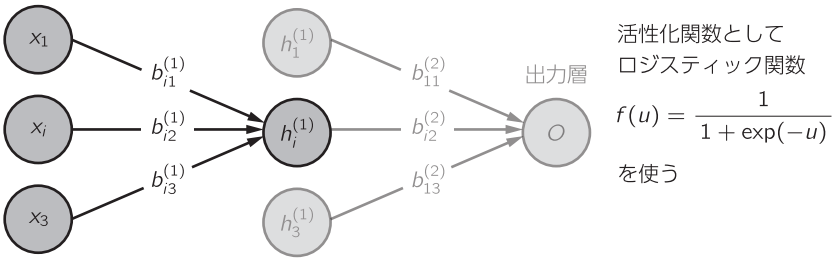

図 8.16　中間層第 1 層のユニット i の値 (出力) を決める式の例

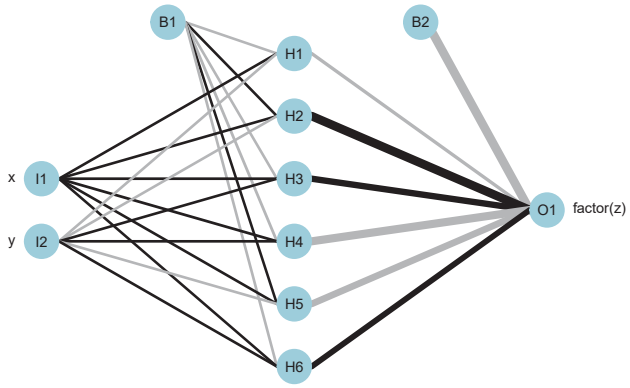

図 8.17　サンプルデータに対するニューラルネットのあてはめモデルの可視化

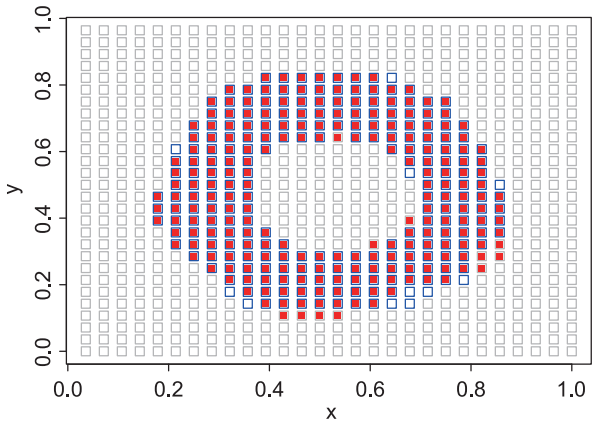

図 8.18　`maxit=2000` を追加したニューラルネットを用いた判別結果 (ユニット数 6)

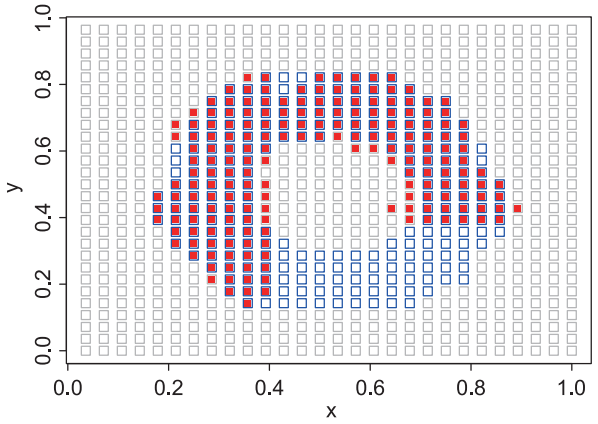

図 8.19　`maxit=100` の場合のニューラルネットを用いた判別結果 (ユニット数 6)

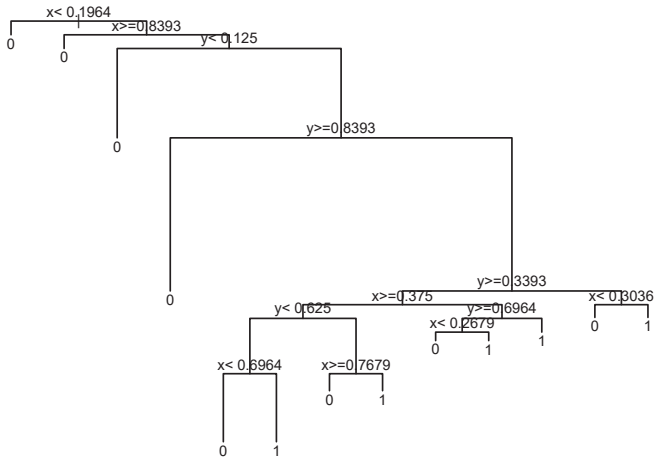

図 8.20　サンプルデータに対する決定木あてはめの樹木図

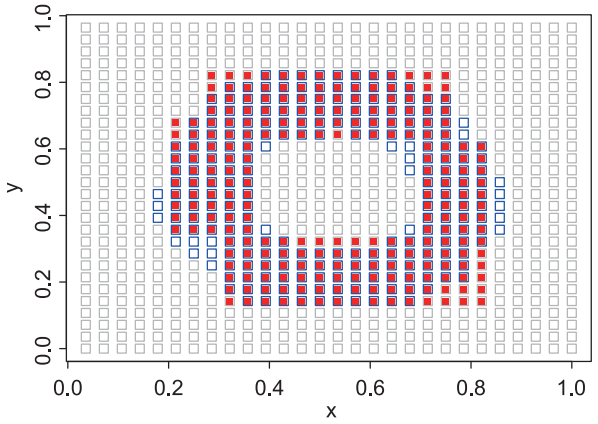

図 8.21　サンプルデータに対する決定木を用いた判別結果

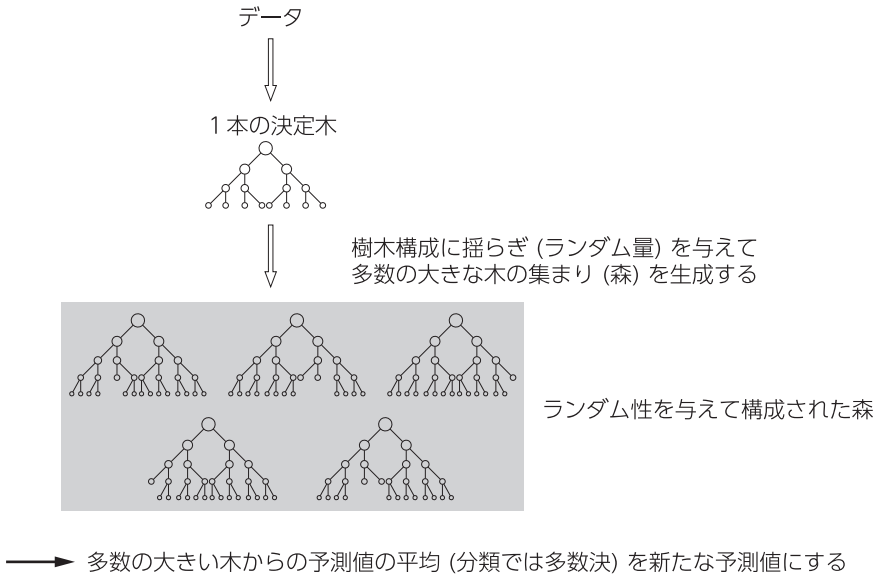

図 8.22　ランダムフォレストの構成方法の概略

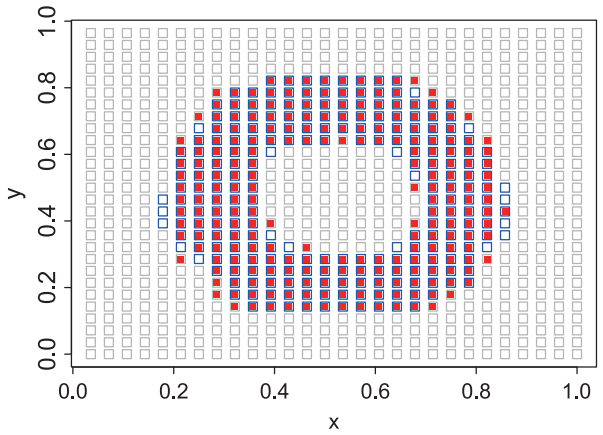

図 8.23　サンプルデータに対するランダムフォレストを用いた判別結果

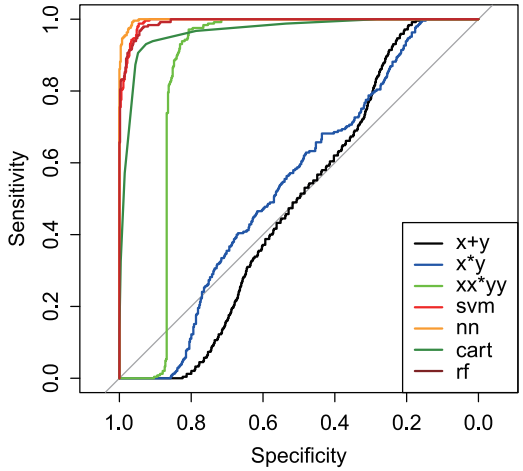

図 8.24　6 つのあてはめ結果 (ロジスティック回帰，SVM，NN，CART，RF) による ROC 曲線の比較

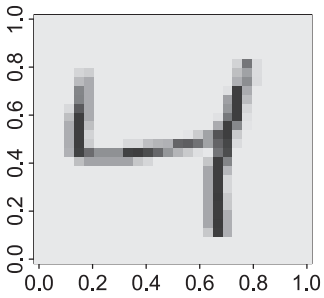

図 8.25　`image` 関数による mnist2000.csv データの 3 行目の表示

予測：未来のデータに対して，　結果 (目的変数) を当てること
予測モデルの作成で用いたデータを**訓練** (**学習**) **データ**という

予測モデル $f(x)$ の例

入力値
説明変数 x　⟹　$a + b_1X_1 + b_2X_2$
ロジスティック回帰での予測
決定木，　RFによる予測
機械学習 (NN, SVM) による予測　⟹　予測値 $f(x)$

興味ある説明変数 x^* を用いて (未来の) 予測値 $f(x^*)$ を得る

図 8.26　教師あり機械学習における予測/予測モデルとは

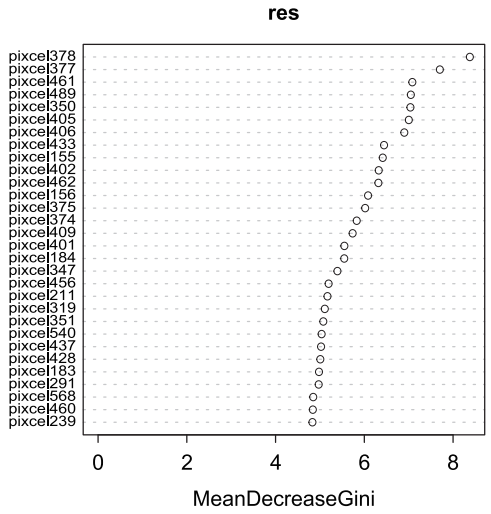

図 8.27　RF の変数重要度プロット

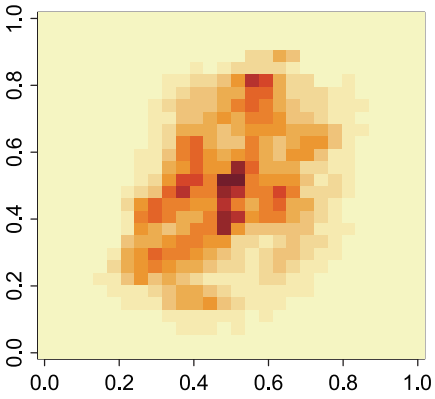

図 8.28　RF の変数重要度のある 2 次元プロット

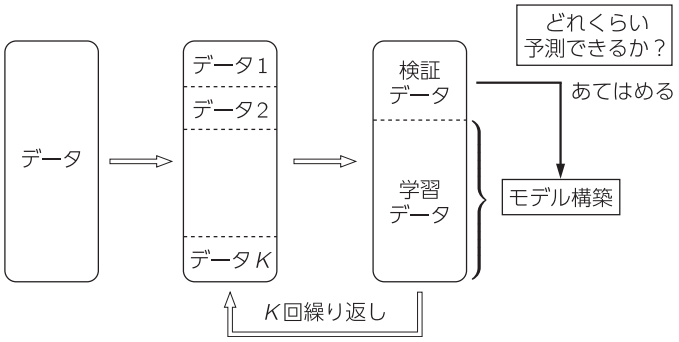

図 8.29　クロスバリデーションの概略

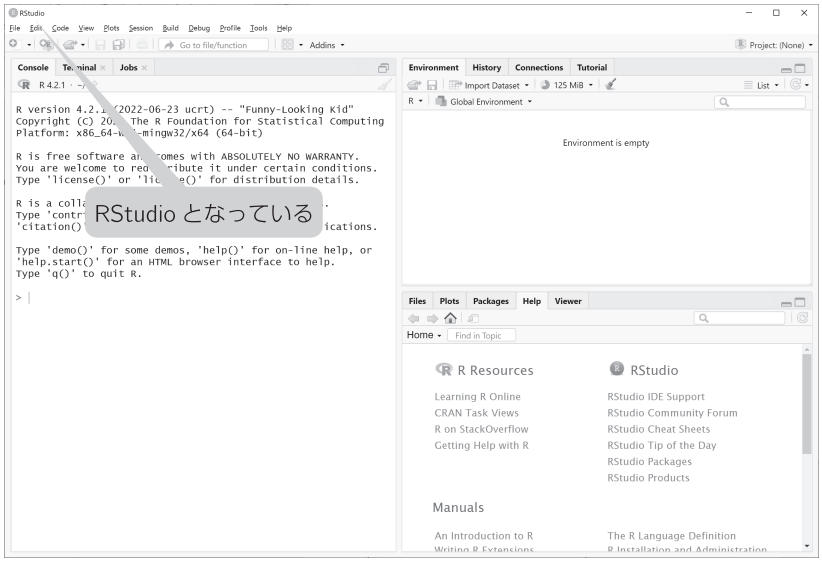

図 P1.1　RStudio を起動したところ

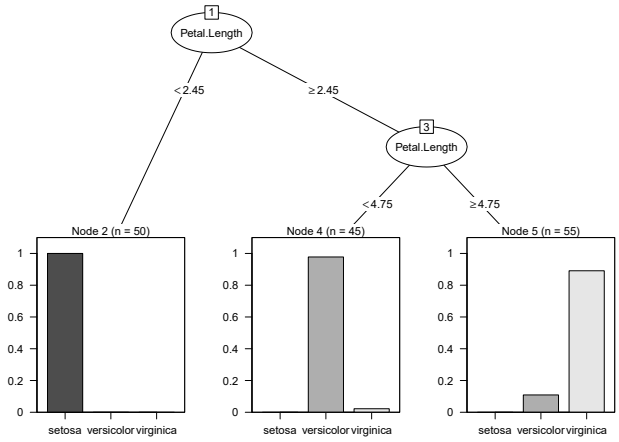

図 P1.2　決定木によるデータ分析の結果

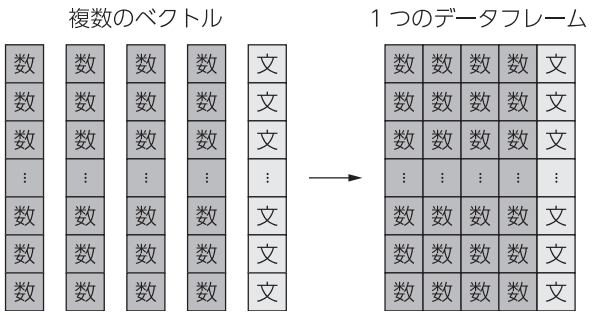

図 P2.1　ベクトルとデータフレームの関係

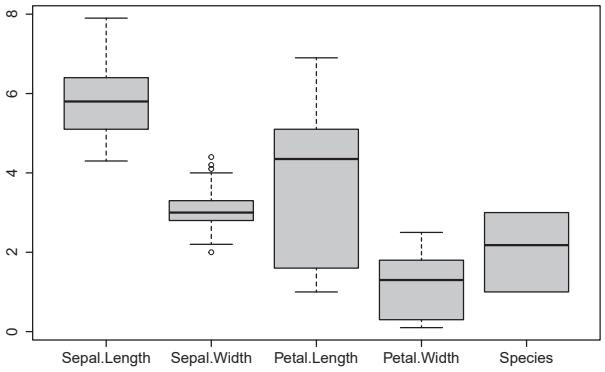

図 P2.2　iris データの箱ひげ図

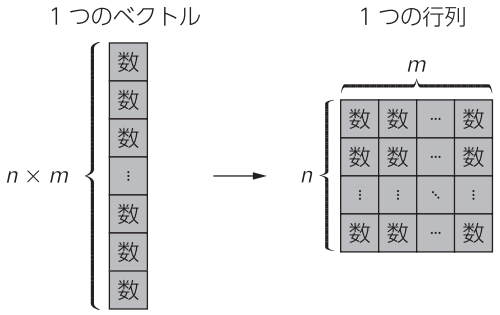

図 P2.3　ベクトルと行列の関係

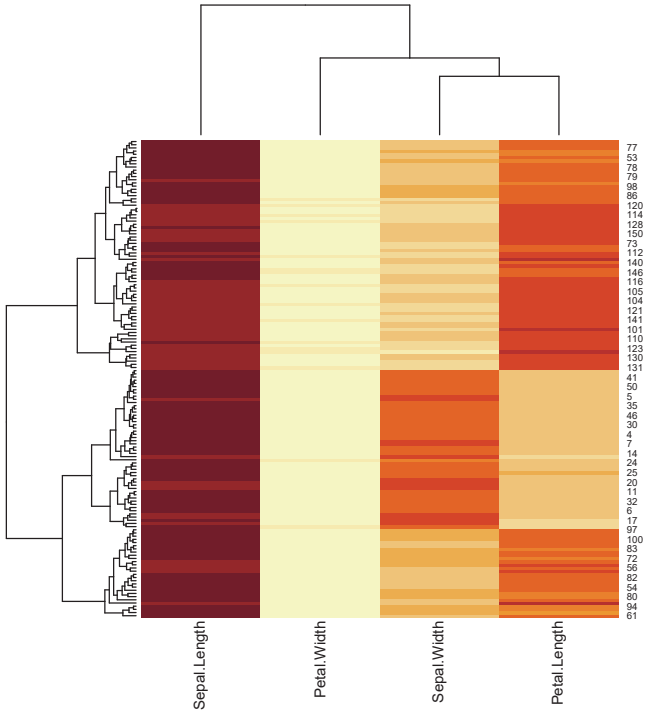

図 P2.4　iris データのヒートマップ

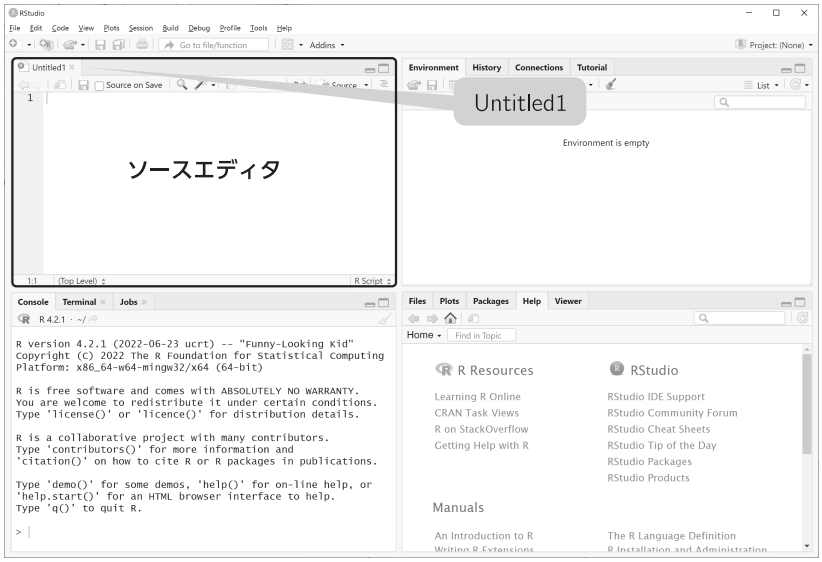

図 P3.1　RStudio のソースエディタ

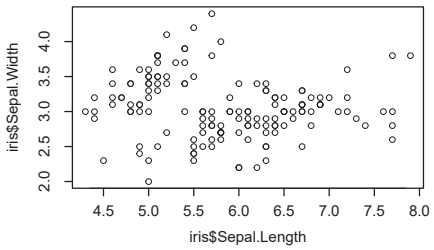

図 P4.1　散布図の例

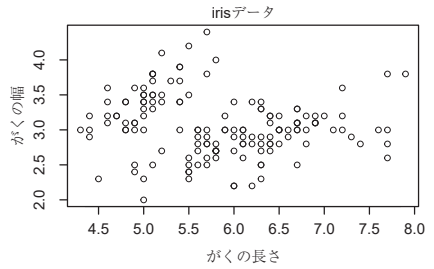

図 P4.2　ラベルとタイトルを日本語で入力した散布図

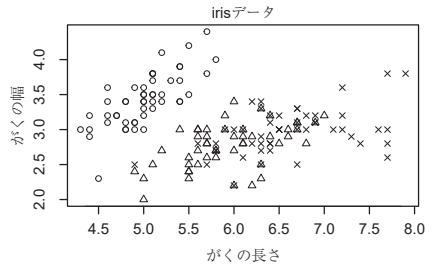

図 P4.3　アヤメの種類ごとに点の形状を変えて作成した散布図

図 P4.4

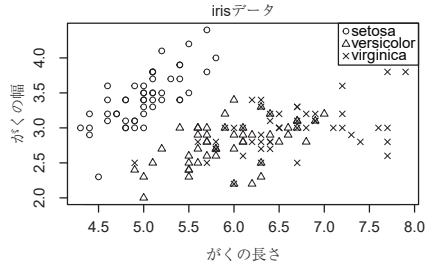

図 P4.5　凡例を追加した散布図

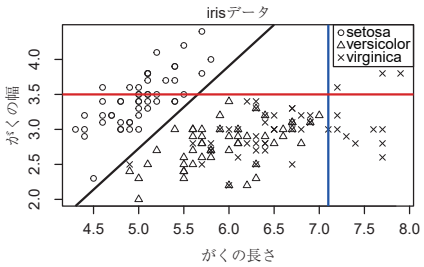

図 P4.6　直線を追加した散布図

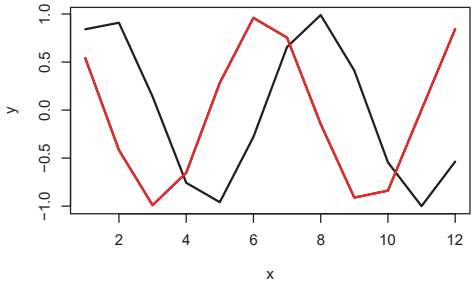

図 P4.7　重ね描きした折れ線グラフ

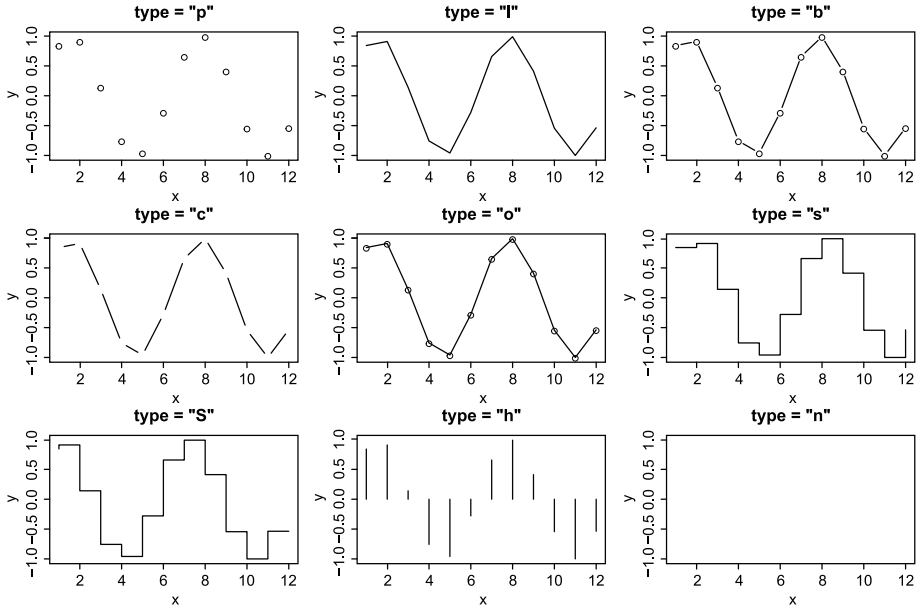

図 P4.8　type オプションによるグラフの表示方法の違い

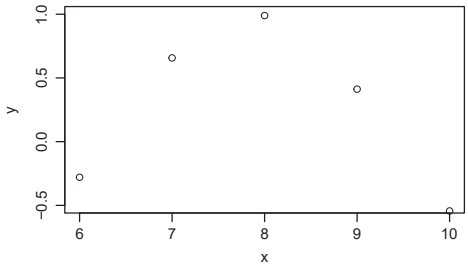

図 P4.9　*x* 座標と *y* 座標の表示範囲を指定したグラフ

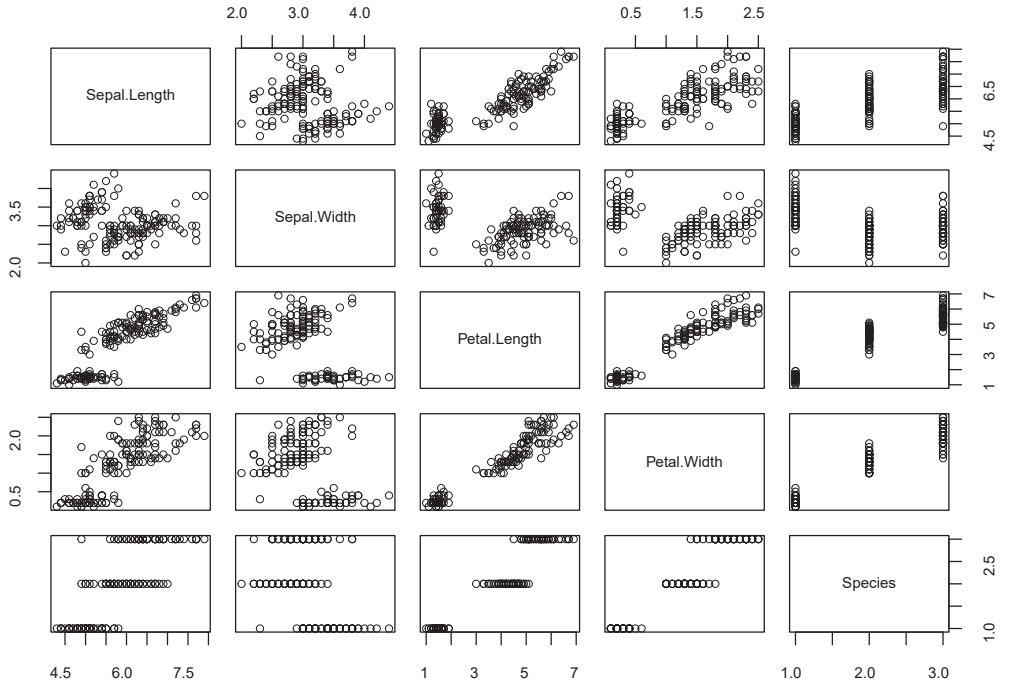

図 P4.10　散布図行列

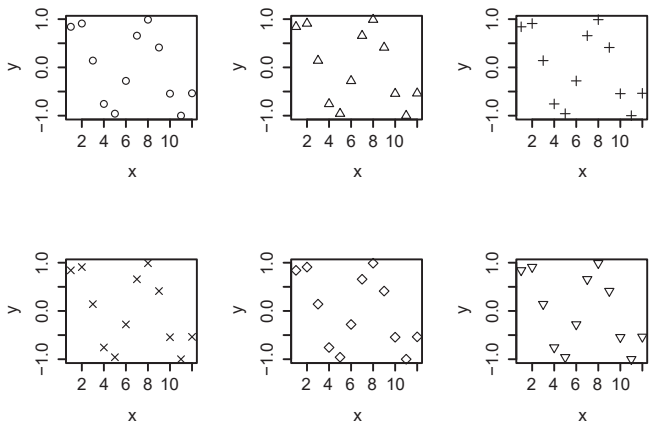

図 P4.11　`mfrow` パラメータを用いた複数グラフのプロット

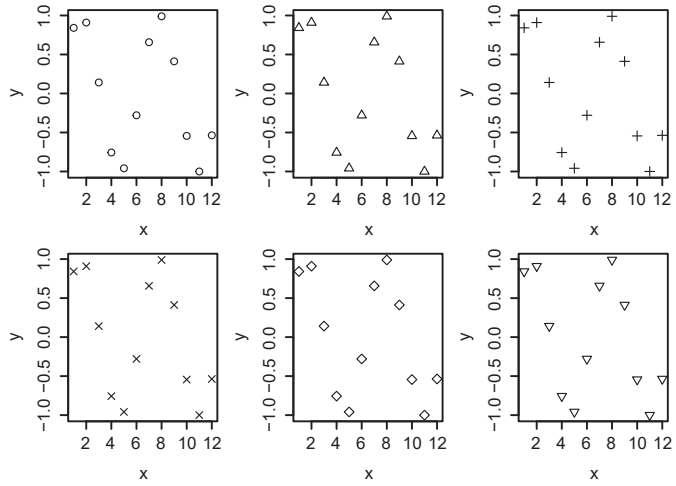

図 P4.12　表示間隔を調整した複数グラフのプロット

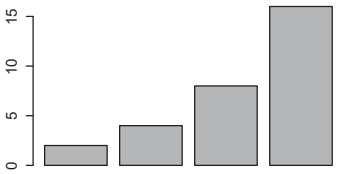

図 P4.13　棒グラフの例

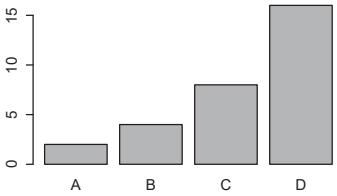

図 P4.14　ラベルを追加した棒グラフ

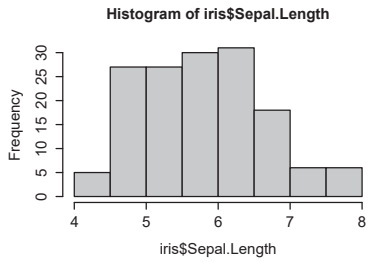

図 P4.15　`iris` データのがくの長さのヒストグラム

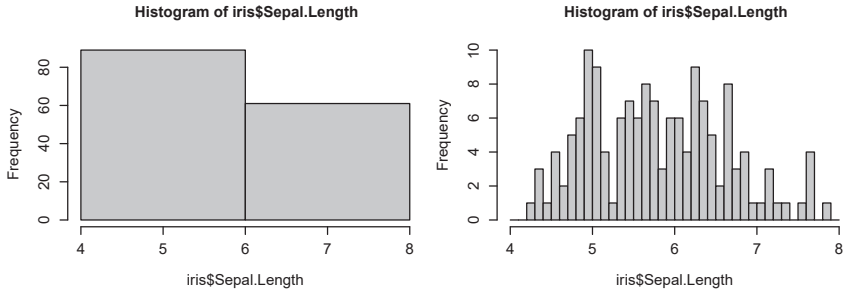

図 P4.16　区間の分け方によってみえ方が異なるヒストグラムの例

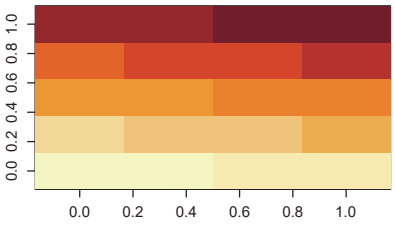

図 P4.17　数値行列の `image` 関数による可視化

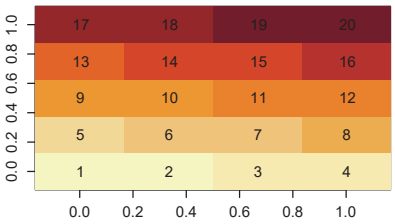

図 P4.18　`text` 関数による数字の表示

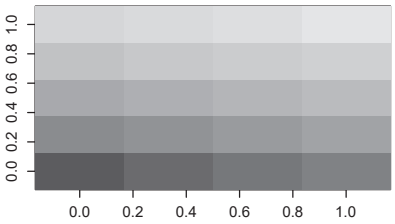

図 P4.19　`image` 関数で表示された図をグレースケールに変更

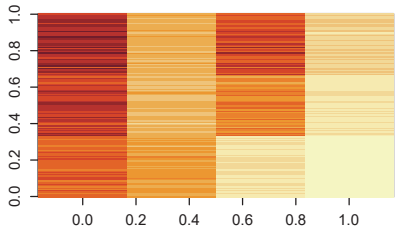

図 P4.20　`iris` データの `image` 関数による可視化

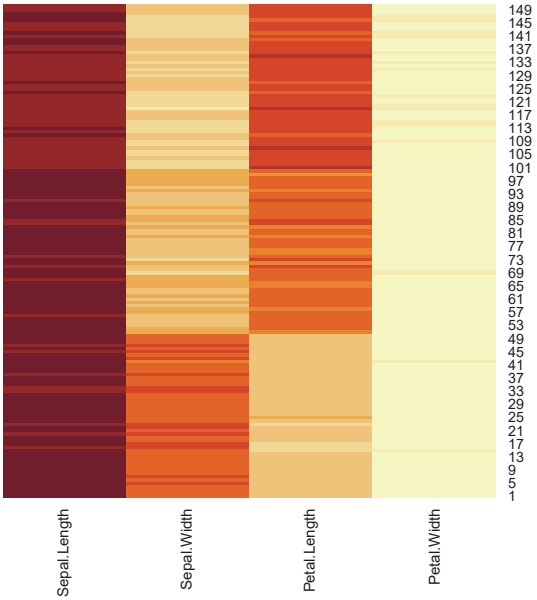

図 P4.21　`iris` データの `heatmap` 関数による可視化 (行に対する標準化)

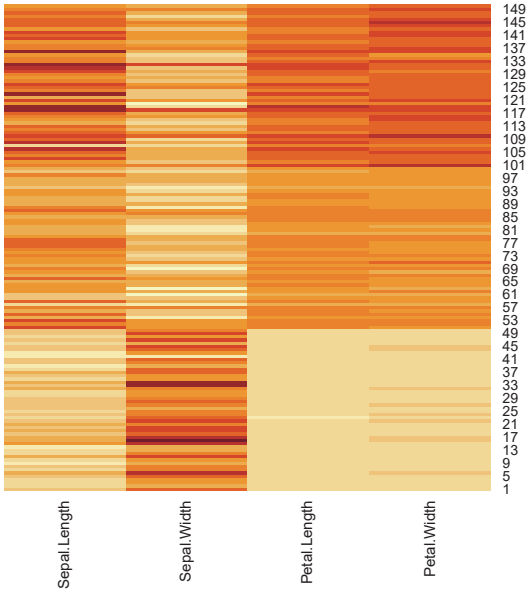

図 P4.22　`iris` データの `heatmap` 関数による可視化 (列に対する標準化)