心理学のためのモデル解析入門

一般線形モデルから多変量解析モデルまで

片所　強

2011 年 10 月 8 日

目次

序説

統計と心理学

拙者が数学やらプログラミングといった類の本を読んでいると、まだそれほど親しくない初対面に近い友人たちは「大学時代の学部は何学科なのか」と尋ねてくる。それに続いて心理学科に在籍していたことをいうと、ほとんどの人が首を傾げてしまう。

元々、多くの学問の研究内容というのは、世間一般から見て未知の領域であることがほとんどである。例えば、我々が義務教育の過程で受ける自然科学一般の領域に含まれる、数学、物理学、化学、生物学といったものでさえも、大学でどのようなことをするのかを知る人など割合でいったらたかが知れているだろう（なんとなく想像はつくかもしれないが）。

心理学という学問の場合、これに含まれる領域があまりにも広すぎて、およそ心理学を学んだことのない者からは想像外のことをやっていることもしばしばある。どうも今日の世間では「心理学＝カウンセリング」というイメージが定着しているようである。しかし、古典的に心理学が得意とする方法論は実験である。実験というのは、ヒトを対象に行うことも多いが、もっと原始的にマウスやハトを用いることもあり、イメージとしては医学や薬学でやっているであろうことをしているのである。

ところがコンピュータ（パソコン, personal computer）の発展と普及に伴ったことも相まって、最近では実験はもとより調査という大規模なデータを得るような機会が圧倒的に増えている。心理学の分野で、卒業論文や修士、博士論文を初めとして、各々の学会が発刊している学会誌に載せる学会論文のなかで調査によって 200 人以上といった大きな数のデータが取り扱われているのである（これに対して実験で得られるデータ数は多くても 20 個とかである）。

このように、心理学ではいずれにしても「データ」をとって、それを「解析」しなければならないのである。ここで「解析」という作業に一役かってくれるのが統計学である。統計学は数学の確率論をベースとして発展していった学問領域の 1 つだが、中学校や高校でやったようなサイコロやコインを投げたり、袋から赤玉と白玉を取り出すというような問題とは少し違う。心理学の中でも数式まみれの生活を送らねばならない分野もあるが、それは少数派にすぎず、多くは分析手法の実務的な意味を理解して、計算はコンピュータに任せれば済んでしまう。

実際、心理学科へと進学する生徒の多くは文系であるので、はなから「セールスお断り」

といわんばかりの勢いで「数式お断り」という姿勢をとってしまうようである。だから、心理学を学ぶものを対象に教科書を書く場合、なるべく数式を使わずに書く必要があるわけである。

しかし、だからといって全く数式なしに、かつ理論的に統計学の諸手法を理解することは難しい。そこで本書では数式はそれほど多く用いない代わりに、モデル式といって、分析者が扱う問題をやや数学的な表現で表したものを用いる（これについて詳しくは後述する）。また、R という無料の統計解析システムを利用して、実際にデータ解析を行いながら「体で覚える」ような内容となっている。

本書における用語と表記

さて本書では「モデル」という用語を多用しているが、これについて少し説明しておくことにしよう。統計学においてモデルという用語を用いる場合、大きく 2 つの意味で使われることが多い。1 つは数学的な意味で使われる場合である。例えば、統計学には回帰分析という手法があるが、これは**線形モデル**と称されるものの 1 つで、数式で表現すると次のように書ける。

$$Y = \alpha + \beta X$$

一方で現象を記述するために**モデル式**として表現される場合がある。この場合のモデルという用語の意味は、前述したような数学的（統計学的）な意味とは若干異なる。例えば、家庭での勉強時間が長いほど学期末テストの総合点が高くなるという現象を説明したいとする。そのとき、勉強時間（$hormework$）というデータと総合点（$total.scores$）というデータが得られることになるだろう。それで「家庭での勉強時間が長いほど学期末テストの総合点が高くなる」という現象をモデル式で記述すると次のように書ける。

$$total.scores = homework$$

これらは一見すると全く違ったものに見えるが、根本的には同じ意味を持っている。ただ、後者は分析者の取り扱う問題（理論的に記述しようとしている現象）を、より現実の場面に適用した形で表現しているに過ぎない。

また、本書ではモデル式に書かれる（分析に使われる）変数をイタリック: *Italic* で、R の出力結果を文章中で説明する場合にはタイプライター: `typewriter` で表記を統一している。

第 1 章

モデル解析という考え方

1.1 測定するとはどういうことか

心理学では統計学を道具として扱う。なぜなら、自然科学としての心理学において、研究とは (1) 仮説を立て、(2) 実験、あるいは調査の計画を練り、(3) その計画に基づいてデータを取ることになる。そして、(4) 実験や調査によって得られたデータを解析する。さらに、(5) その解析結果を解釈するという 5 段階の課程を経るからである。

これらすべての段階において、統計学の基礎的な知識が必要となる。統計学というと、もっぱら (3) のデータ解析のみがそれであると思われがちである。しかし、適切なデータ解析というものは、そもそも正しく計画された実験や調査によって得られた、いわゆる質の高いデータが得られていなければ成り立たない。

ここで質の高いデータとは、研究者が測りたいと考えているものを忠実に表された情報のことである (心理学では主に心の働きが測定の対象となるだろう)。これは測定尺度の妥当性の問題として考えられることである。例えば、男性らしさや女性らしさといったことを、身長や体重で説明してみようとする（生物学的には、体型は性別を区別するための重要な指標となるだろう)。このとき、身長を測定するためには (1) 適切な尺度 (測定器具) の作成、(2) 精度の高い測定技術、といった 2 つの技術的な問題が発生する。

まず第一に身長を単位 cm で測定する場合、単位が cm で表された測定器具が必要となるだろう。また、長さを測る測定器具であっても、それが in($1\ \text{in} = 2.54\text{cm}$) という単位で表された測定器具であれば、正確な cm 単位の身長を測定するのは難しいかもしれない。それは、単位を変換する際に生じる誤差があるためである。あるいは、もっと的外れな例をあげれば、身長を測るのにストップウォッチを用意してもまるで役に立たないだろう。これは身長を測るのに適さない測定器具なのである。

第二に、いかに適切な尺度を用いたとしても、あるものを正確に測るにはある種の測定技術が必要となる場合がある。測定技術を違う言葉で言い換えると、それは測定器具の使い方と測定方法の正しい手順を示したマニュアルに従う技術といえる。例えば、我々はアナログ式でもデジタル式であっても、時計の読み方を小学校で習っているのでほぼ正確に時計を読

むことができる。しかし、過去にに時計の読み方を教わっていなければ、時計という測定器具を用いて時間を測定することはできないだろう。

一方で時計の読み方を知っているからといって、正確な測定ができるとは限らない。例えば、100m 走のタイムを計測する場合を考えてみよう。もしかしたら、ある人は「スタートの合図 (火薬銃の「バンッ」という音) が聞こえたら、ストップウォッチのボタンを押し、ゴールラインへ体の 1 部が触れた (超えた) 瞬間にボタンを押す」という測定方法の手順に従うかもしれない。しかし、これは正しくない。正確にタイムを計測するためには、スタート合図の音が聞こえた瞬間にボタンを押すのではなく、火薬銃の煙が見えた瞬間に押さなければならないのである (それは音の速さより光の速さの方がより速いからである)。

このように、いかに正確な測定器具を用いたとしても、それの正しい使い方と測定の手順 (方法) を理解して実行できなければ、それは質の低いデータとなってしまうのである。こうした物理的に観測可能なものを測るのでさえも、質の高いデータを取るのが相当に面倒なことであると分かるだろう。これがヒトの心といったように、そもそも直接的に観測不可能なものを測ろうとしているのが、心理学なのである。だからこそ、心理学においてデータの測定には他の学問よりも妥当性について考慮しなければならない。つまり、測定器具と測定技術の限界を知れということである。

1.2　実験と調査の計画

さて、心理学においてデータを取るということが大変なのだろうということは想像できたかもしれない。ではどうしたら、より質の高いデータを取ることができるかを考えてみよう。これが実験と調査の計画という段階で考えられるべきことである。

まずデータを取る際に重要なことは、自分が最終的に主張したいことは何かと自問自答することである。つまり、良い仮説を立てることである。良い仮説とは反証可能な仮説のことである。例えば、以下の 2 つの仮説を見てみよう。

1. ネス湖にはネッシーが存在する
2. ネス湖にはネッシーは存在しない

まず仮説 1 が反証可能かどうか考えてみる。ある期間内にネス湖をくまなく探したが、結局ネッシーの存在を確認することはできなかったとする。ここで、もしかしたらその調査者は「ネス湖をくまなく探したが、ネッシーは見つけられなかったので、仮説 1 は正しかった」と主張するかもしれない。

しかし、これでは仮説を反証したとはいえない。なぜなら、この調査で分かったことは「ある期間、ある観測地においてネッシーの存在を確認できなかった」というだけのことだからである。もしかしたら、たまたま調査した場所にはネッシーが居なかったのかもしれない。そう考えると、いくらネス湖を調査しても仮説を棄却することはできないだろう。

これに対して仮説 2 はどうだろうか。これについては、たとえ 1 匹でもネッシーの存在を

確認することができれば、仮説を棄却することができる。つまり、理論的には反証することが可能なわけである。見方を変えると、仮説 2 は棄却も採択もできるが、仮説 1 は採択することしかできない。だから、自分で仮説を立てるときはこのことを気に留めておく必要がある（科学ではこれを反証可能性という）。

1.3 モデル式による表現

順序的には、ある現象について科学者は疑問を持ち、そこからある仮説を立てることになる。次いで、対象となるものが測定可能なものか、可能ならどのような測定器具でどのような方法を用いて測るかを考える。そして、実験や調査の制約条件の下で最も効率よくデータを取ることを考えることになる。この 1 連の操作が実験計画、あるいは調査計画というものである。

実験計画の段階で、自分が主張したい仮説を立てたのであれば、それをモデル式で表現する必要がある。なぜならば、モデル式は扱う問題をより単純に、かつ明確に表現したものだからである。また、データ解析の段階でどのようなモデルを解析するのかを明らかにする意味でも、仮説をモデル式で表現することは重要である。

例えば、ある研究者は「うつ傾向を表す尺度得点が高い被験者は、内向性傾向を表す尺度得点も高くなる」という主張がしたいとする。注意すべきは、これは主張したいことであって仮説ではないということである (主張＝仮説というわけではない)。この問題を換言すると「うつの尺度得点と内向性の尺度得点には相関関係がある」ということができるだろう。しかし、これは前述した仮説 1 と同じで、反証することができない仮説である。したがって、良い仮説ならば「うつの尺度得点と内向性の尺度得点には相関関係がない」と書くべきである。

1. うつの尺度得点と内向性の尺度得点には相関関係がある (悪い仮説)
2. うつの尺度得点と内向性の尺度得点には相関関係がない (良い仮説)

そしてこれは式 1.1 のように表現することができる。モデル式において、等号 (=) は数学でいうところの「等しい」という意味ではない。これは「右辺は左辺を説明する」という意味であり、左辺に置かれる変数を**応答変数**といい、右辺に置かれる変数を**説明変数**という。

$$\text{うつの尺度得点} = \text{内向性の尺度得点} \tag{1.1}$$

統計学の理論に基づいてデータ解析を行う場合、解析の対象となる問題は必ずモデル式で表すことができる。逆にいえば、モデル式で表現できない問題は解析することができないということである。つまり、実験計画の段階でモデル式を考えることによって、データを取った後に適切なデータ解析法を選べないという問題を避けることができる。心理学の分野ではよく見受けられるが、とりあえずデータを取ってみて、その後に使えそうな分析法を探すというのは適切ではない。

1.4　連続型とカテゴリカル型

モデル式を書く場合に変数の型を考えなければならない。そもそも変数とは、データが格納されている箱のことであり、そこに格納されているデータは必ず比尺度、間隔尺度、順序尺度、名義尺度といった 4 つの**尺度水準**を満たしている。

データは大きく**連続型**と**カテゴリカル型**の 2 つに区別して考えることができる。連続型は離散的ではない値のことで、分かりやすくいえば小数点を伴うデータは全て連続型であるといえる。長さ (cm, in, ft)、質量 (kg, lb, ct)、温度とエネルギー (℃, K, J, cal) といったものは全て連続型である。違う言い方をすれば、**間隔尺度**以上の尺度水準を満たすデータは連続型である。ちなみに、人数や金額は小数点を伴うことはないが、間隔尺度の尺度水準を満たしているので、場合によっては連続型として扱っても問題ない。

一方でカテゴリカル型は尺度水準として**名義尺度**である場合と順序尺度である場合の 2 つが含まれる。そもそもカテゴリカル (categorical) とは category の形容詞で、これには「分類上の区分、種類、範疇」という意味がある。例えば、性別は男性と女性という 2 つの**属性** (**水準**) をもったカテゴリカル型変数である。都道府県は 47 つの属性をもったカテゴリカル型変数である。少し変わった例をあげれば、郵便番号や電話番号もカテゴリカル型である。

それでは**順序尺度**である場合とは、どのような例があげられるだろうか。例えば、年齢層 (3 歳未満、3 歳〜15 歳、16 歳〜59 歳、60 歳以上) というのは 4 つの属性をもったカテゴリカル型の変数だが、これには順序関係を持たせることができる。3 歳未満は 3 歳〜15 歳の区分よりも低い、あるいは 60 歳以上は 16 歳〜59 歳という区分よりも高いといったような順序関係があるわけである。学年 (中学 1 年、2 年、3 年) なども 3 つの属性をもったカテゴリカル型であろう。

こういった順序尺度は名義尺度として扱うこともできる。原則としてより上位の尺度水準を満たすデータは、その尺度水準よりも下位の尺度水準として扱っても問題ない。ただし、その際には重要な情報を削ぐことにもなるので、安易に下位の尺度水準として扱うのは好ましいことではない。

1.5　一般線形モデルと一般化線形モデル

モデル式において、左辺に置かれる変数のことを応答変数、右辺に置かれる変数を説明変数であると述べた。そして変数には連続型とカテゴリカル型とに区別して考えることができるということも述べた。なぜこのようなことを考えなければならないのかというと、実は以下の 4 項目について考えたとき、その条件によって解析される**統計モデル**が変わるからである。

1. 目的変数の数はいくつか
2. 目的変数の型は何か

3. 説明変数の数はいくつか
4. 説明変数の型は何か

多くの場合において項目 1 については考えなくてよいだろう。なぜなら、以降に紹介する統計モデルのほとんどは目的変数が 1 つの場合だからである。目的変数については連続型かカテゴリカル型か (上記項目 2) について確認しておくことのほうが重要だろう。

なぜなら、説明変数はいくつあっても構わないし、連続型とカテゴリカル型が混在していても構わない。ただし、統計モデルはなるべく単純であることが好ましいという原則は常に気に留めておくべきである。

例えば、カテゴリカル型のみを扱う場合、その数は多くても 3 つまでにしておいたほうがよい。連続型のみを扱う場合（重回帰）であっても、多くても 7〜10 個程度に留めておくほうがよいかもしれない。両者を混在させる場合（共分散分析）には連続型が 1 つ、カテゴリカル型が 1 つとするくらい単純な方がよいかもしれない (特に初心者にとっては)。

表 1.1 に代表的な統計モデルをまとめておいた。統計モデルは一般線形モデルと一般化線形モデルとにわけられる。一般線形モデルは目的変数が正規分布に従うと仮定されることを前提とした統計モデルである。この一般線形モデルを拡張したものが一般化線形モデルに含まれるモデル群であり、目的変数が正規分布ではない場合に用いられるものである。例えば、誤差がポアソン分布、二項分布、多項分布などに従うと仮定される場合に適用することが可能である。

表 1.1　統計モデル一覧

モデル式	モデル名	説明変数の型	目的変数の型
$Y = A$	1 要因分散分析	カテゴリカル	連続
$Y = A + B$	2 要因分散分析	カテゴリカル	連続
$Y = X$	単回帰分析	連続	連続
$Y = X_1 + X_2 + ... + X_n$	重回帰分析	連続	連続
$Y = A + X$	共分散分析	混在	連続
$Y_1, Y_2, ..., Y_n = A$	多変量分散分析	連続	連続
$Y = A + B$	ポアソン回帰	カテゴリカル	計数データ
$Y = X$	ロジスティック回帰	連続	比率データ
$Y = X_1 + X_2 + ... + X_n$	2 値ロジット	何でもよい	2 値データ
$Y = X_1 + X_2 + ... + X_n$	名義ロジット	何でもよい	名義データ
$Y = X_1 + X_2 + ... + X_n$	順序ロジット	何でもよい	順序データ

第 2 章

古典的な検定

2.1 t 検定

心理学において最もよく使われる検定手法の 1 つとして t 検定があげられる。**t 検定**とは、より具体的にいえば**独立 2 標本の平均値の差の検定**のことである。t 検定という名称は **t 分布**を利用した検定の総称であるので、どのような場合においても t 検定と記すのは誤解を招くことになりかねない。

ここでは最も単純な例をあげて、t 検定がどのようなものであるかを紹介しよう。表 2.1 は男性の身長データと女性の身長データをまとめたものである。分析者の疑問は「男性身長の平均値と女性身長の平均値は異なるか」ということである。

表 2.1　男性と女性の身長データ

男性	女性
175	160
180	155
169	167
170	170
173	158

2.1.1 帰無仮説と対立仮説

統計学において仮説は**帰無仮説**と**対立仮説**を立てることになる。帰無仮説は H_0 で表され、多くの場合は棄却されることが望まれるだろう。それに対して対立仮説は H_1 で表される。いうまでもないことだが、帰無仮説が棄却されるということは対立仮説を採択することである。逆に帰無仮説が採択されれば対立仮説は棄却されることになる。

- H_0: 男性の母平均と女性の母平均は等しい

- H_1: 男性の母平均と女性の母平均は異なる

これはより数学的に表記することもできる。男性の母平均を μ_1 とし、女性の母平均を μ_2 として帰無仮説と対立仮説を改めて書くと次のようになる。

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

2.1.2 母集団と標本

ところで、ここに**母平均**という新たな用語がでてきた。これは母集団の平均値のことを指しており、統計学では常に母集団と標本を比較して考えることになる。我々は母集団がもつ値 (母平均値や母分散) を決して知ることはできない。したがって、母集団の代表となる標本から得られた値 (標本平均や標本分散) を推定するしかないということになる。多くの参考書で単に平均と記している場合、まず間違いなく標本平均のことを指している。

話が少しずれるが、母集団とは哲学者プラトンのいうイデアのことである。これは我々が認識している世界は不完全なもので、イデアという完全なる世界が別に存在するという考え方である。実際、厳密にいえばある線分の長さを正確に知ることはできない。例えば 30cm の線分を定規で測ろうとしても、限りなく小さな単位まで正確に測ろうとすれば必ず測定誤差が生じるだろう。分かりやすくいえば、10 回繰り返して測定すれば、10 回とも異なる測定値が得られるということである。極端にいえばナノメートルという単位まで測ろうとした場合には、明らかな**測定誤差**が生じるだろう。もちろん、これは理論的 (仮定的) な話である。

こうした場合、我々はいかようにしても真の値 (30cm) を測定することができない。そこで得られている測定値 (標本値) から真の値 (母集団値) を推定するわけである。ここであげた例は非常に極端な例であるが、現実的に日本に在住する人全ての身長を把握しようとしても不可能である。なぜならば、1 つは母集団は流動的であるので、観測期間中に母集団の値は変動してしまうからである。また、別のもう 1 つの理由として、それだけ多くの値を得ようとすると測定誤差が生じる可能性が高いということがあげられる。

2.1.3 R

これ以降の解析例はすべて R という無料の統計ソフトを用いて行っている。R についての基本的な情報は次の web サイトより得ることができる。

- http://www.okada.jp.org/RWiki/

また著者の管理する web サイトでも、R の操作方法について紹介しているので参照してもらいたい。

- http://homepage2.nifty.com/nandemoarchive/sas_r_excel/mokuji_r.htm

2.1.4 R による解析例

本書では R の操作方法について詳しく解説していないが、全ての解析例について R による解析手順を示してある。実際に R を用いて例題を解きながら本書を読み進めていくと、より効果的な勉強ができるであろう。また、紙面の都合上、R の出力結果を一部的に省略して載せてあることもここで断っておく。

```
> male <- c(175, 180, 169, 170, 173)
> female <- c(160, 155, 167, 170, 158)

> t.test(male, female, var.equal=FALSE)

        Welch Two Sample t-test

t = 3.3243, df = 7.156, p-value = 0.01229
```

分析者がとりあえず見るべき値は p 値 (p-value) であろう。なぜならば、p 値を見ることによって直接的に帰無仮説を棄却するか採択するかを判断することができるからである。この場合、$p = 0.01229$ となっているので 5% の有意水準のもとでは帰無仮説を棄却することになる。

2.1.5 p 値と有意水準

p 値は「帰無仮説が正しいという仮定の下で、偶然に検定統計量 (この場合は t 値) よりも大きな値が得られる確率」のことである。$p = 0.01229$ という値は偶然に検定統計量 ($t = 3.3243$) よりも大きい値が得られる確率が非常に小さいことを意味しているのである。すなわち、実際に検定を行う際には帰無仮説を棄却するための証拠力として考えればよい。

別の言い方をすれば、p 値は検定を 100 回繰り返し行ったら、そのうち n 回は誤った結果が得られる確率ということができる。だから、$p = 0.01229$ という値は「100 回のうち 1 回は誤った結果を得てしまう可能性がある」ということを意味している。

これに関連して**有意水準**について正しく理解しておくことが重要となる。p 値が 100 回のうち何回失敗するかという値であるならば、何回の失敗なら許容するかという値を検定を行う前に設定しておく必要があるだろう。それが有意水準と呼ばれるものである。

心理学では慣例的に有意水準を 5% と設定することが多い。つまり、これは 100 回のうち 5 回の失敗なら許容しようという基準を設けているわけである。薬の開発のような、臨床試

験の場面ではしばしば 1%、あるいは 0.1% という、より厳しい有意水準が設定されることがある。いずれも、数学的な理論を基に設定されるものではなく、あくまでも慣例的であるということに注意すべきだろう。また、有意水準は α で表され、5% の有意水準を $\alpha = 0.05$ と書き記すことが多い。

2.1.6　t 分布: t 値と p 値の関係

まずは図 2.1 を見てみよう。これは自由度 $(df, degrees of freedom)$7.156 の t 分布である。t 分布は正規分布のように左右対称の鐘形の曲線を描くようになっている。分布に関しては他の統計学の教科書を参照してもらいたい。

先ほどの解析例で得られた $t = 3.3243$ という値は、図 2.1 の横軸の何処に位置するかというものである。そして p 値とは、その値を境界として右側の面積である (図の灰色に塗りつぶされた部分)。

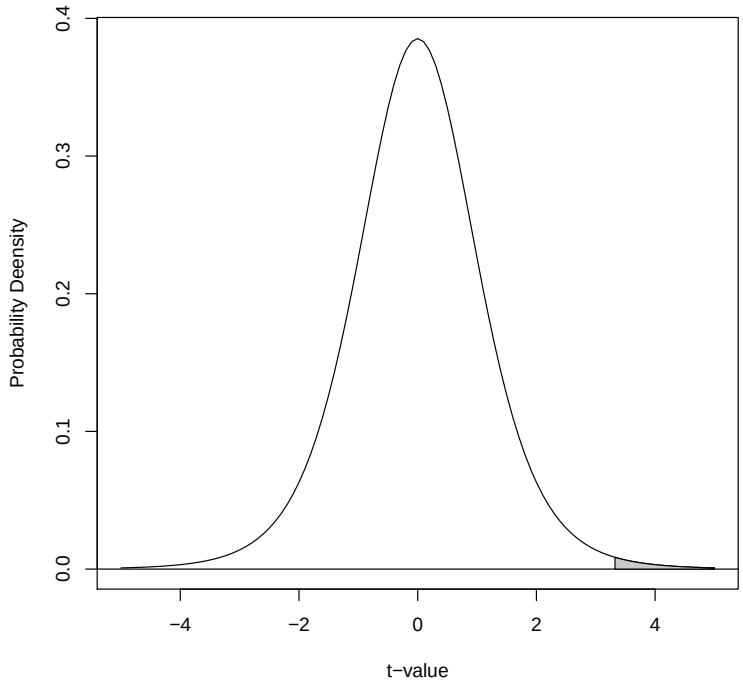


図 2.1　自由度 7.156 の t 分布

R には t 値と自由度を指定すれば、その値を境界として左側の面積 (p 値) が返される `pt()` という関数が用意されている。それを利用して確認してみよう。

```
> 1 - pt(3.3243, df=7.156)
[1] 0.00614507
```

これで図 2.1 の灰色の部分の面積 (p 値) が分かった。ところが、これは先ほどの解析例で得られた $p = 0.01229$ と一致しない。これはどういうことだろうか。結論をいうと。先の解析では**両側検定**を行っているので、0.00614507 という値を 2 倍しなければならない。

```
> (1 - pt(3.3243, df=7.156)) * 2
[1] 0.01229014
```

検定には両側検定と**片側検定**とがあるが、心理学ではもっぱら両側検定が行われる。片側検定は、例えば、「女性より男性の平均の方が大きい」という両群の平均値の大小を比較する際に用いられる。詳しくは他の参考書の解説にゆずることにする。

ここでは分布と両側検定、片側検定について少しだけ触れたが、より詳しい知識を得るためには参考文献 [9] が優れている。

2.2 分散分析

2 つの平均値の比較は t 検定によって行われた。ここでは 3 つ以上の平均値を比較する場合に用いられる**分散分析**について紹介する。ここで、平均値の比較をするのに、なぜ分散分析という名前が出てくるのか不思議に思うかもしれない。これについては 6 章で詳しく説明しているので、そちらを参考にしてもらいたい。

この分散分析という手法は、おそらく心理学の中で最も頻繁に利用される統計学の手法である。それゆえに分散分析で得られる**分散分析表**の読み方を理解しておくことは重要である (しかし、これは 3 章で紹介する分散分析モデルのところで解説する)。

例えば、表 2.2 はある中学校において、各クラス (松、竹、梅) 代表 7 人の学期末テスト 5 教科の総合得点をまとめたものである。ここで分析者が知りたいのは、3 つのクラスそれぞれで平均値に差があると主張できるかどうか、ということである。帰無仮説と対立仮説は次のように立てられる。

- H_0 : 3 クラスの平均値は等しい ($\mu_1 = \mu_2 = \mu_3$)
- H_1 : 3 クラスの平均値は等しくない ($\mu_1 \neq \mu_2 \neq \mu_3$)

それでは、上述の帰無仮説について分散分析を行ってみよう。R には `oneway.test()` という**一元配置分散分析**を行うための関数が用意されている。なお、表 2.2 のようなデータ形式を一元配置といい、これは松、竹、梅という 3 つの属性をもったクラスという要因 ($class$) が 1 つだけ配置されていることを意味する。このことから、一元配置分散分析は 1 **要因の分**

表 2.2　3 クラス代表 7 名の総合得点

松	竹	梅
374	461	356
325	361	285
203	316	437
329	282	315
228	423	415
354	328	381
417	397	324

散分析とも呼ばれる。より正確にいえば 1 要因 3 水準の分散分析といえる。

```
> score <- c(
+ 374, 325, 203, 329, 228, 354, 417,
+ 461, 361, 316, 282, 423, 328, 397,
+ 356, 285, 437, 315, 415, 381, 324)
> group <- rep(1:3, c(7, 7, 7))
> oneway.test(score ~ group, var.equal=FALSE)

        One-way analysis of means (not assuming equal variances)

F = 0.8551, num df = 2.000, denom df = 11.797, p-value = 0.45
```

以上の結果から、$p = 0.45$ という値が得られているので、これは 5% の有意水準の下では帰無仮説は採択されることになる。したがって、各クラスの平均値が異なるとはいえない。

2.2.1　分散分析と F 分布

t 検定では t 分布を利用したが、分散分析では F 分布と呼ばれる分布を利用する。F 分布は 2 つの自由度をもち、今回は $df_1 = 2.000$ と $df_2 = 11.797$ という 2 つの自由度であることが分かる。これは図 2.2 のような分布曲線が描かれることになる。

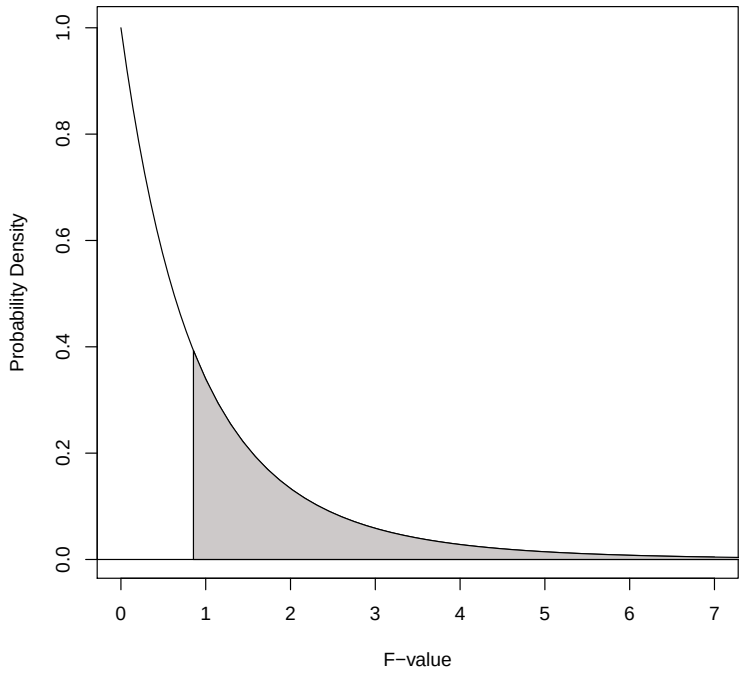


図 2.2 $df_1 = 2.000, df_2 = 11.797$ の F 分布

2.2.2 上側確率と下側確率

ここで、また新たな用語について学ぼう。図 2.2 において、分散分析によって得られた F 値 $(F = 0.8551)$ を境界として右側の部分を灰色で塗りつぶしてある。また、得られた p 値 $(p - value = 0.45)$ という値は、この灰色部分の面積を表している。この考え方は t 検定のときと全く同様であるし、他の検定においても同じである。

検定によって得られた p 値は、まぎれもなく灰色部分の面積を表しているのだが、これは**有意確率**という言い方をすることも覚えておくとよい。そして F 値のような検定統計量のことを**境界値**と表現することもある。この境界値の左側の面積部を特に**下側確率**といい、これに対して右側の面積部を**上側確率**という。

復習になるが、確率密度関数が描く分布曲線の全面積は 1 であるから、R の `pf()` を利用して p 値を求めるためには次のようにする必要がある。ちなみに、「分布曲線の全面積」という表現はおかしなものだが、ここでは便宜的にこのように表現したことを断っておく。正確にいえば、0 から無限大までの区間を積分することで全面積が求められることになり、その面積が 1 になるということである。

```
> 1 - pf(0.8551, 2, 11.797)
[1] 0.4499923
```

下側面積は以下のようにすれば計算できることが分かるだろう。R の各分布についての p 値を求める関数 `pnorm()`, `pchisq()`, `pt()`, `pf()` などは指定した境界値までの面積 (つまり下側確率) を返す仕組みになっている。なお、「返す」というのは、関数が処理を行って、その結果を出力することをこのようにいうのである。

```
> pf(0.8551, 2, 11.797)
[1] 0.5500077
```

ところで、t 検定では「両側検定なので p 値を 2 倍しなければならない」と述べたが、分散分析においてはそのようなことをする必要はない。この点については参考文献 [10] を参照するとよい。

2.3　独立なデータと対応ありのデータ

t 検定の解析に用いたデータセット (表 2.1) と分散分析の解析に用いたデータセット (表 2.2) はいずれも**独立なデータ** (対応なしのデータ) である。これに対して**対応ありのデータ**というものも存在する。

両者の区別は表 2.3 と表 2.4 を見れば明らかである。表のセルに書かれている S は Subject(被験者) の S であり、添え字は被験者番号である。表 2.3 では 15 名の被験者を対象に実験が行われている。一方で表 2.4 は 5 名の被験者を対象に実験が行われている。

独立なデータとは、各条件に配置される被験者に重複がない場合を指す。それに対して対応ありのデータとは、1 人の被験者が異なる条件下に繰り返し参加しているのである。換言すれば、1 人の被験者から繰り返しデータが取られているということである (独立なデータでは、1 人の被験者から得られるデータは 1 つだけである)。

2.4　反復測定と繰返測定

本書で紹介する統計モデルの数々は、独立なデータであることが前提とされて解析されるものである。つまり、対応ありのデータは特殊な場合であると考えてもよいだろう。もちろん、この特殊な場合に対しても適切な解析法は存在する。それが 5 章で紹介する一般線形混合モデルと呼ばれる統計モデルである。

表 2.3　独立なデータ ($n = 15$)

A	B	C
S_1	S_6	S_{11}
S_2	S_7	S_{12}
S_3	S_8	S_{13}
S_4	S_9	S_{14}
S_5	S_{10}	S_{15}

表 2.4　対応ありのデータ ($n = 5$)

A	B	C
S_1	S_1	S_1
S_2	S_2	S_2
S_3	S_3	S_3
S_4	S_4	S_4
S_5	S_5	S_5

また、対応ありのデータはしばしば**繰返測定データ**、もしくは**経時測定データ**と呼ばれている。これはある 1 人の被験者から時間経過に伴って繰り返しデータが取られるという手続きから、このような名称がつけられているのであろう。これに関して、専門家の間でさえも混同されて使われてしまっている**反復測定**という用語について説明しておく。反復測定データとは既に紹介した表 2.3 のような、独立なデータのことである。ここでいう反復とは、フィッシャーの 3 原則である反復 (replication)、無作為化 (randomization)、局所管理 (local control) にあげられている反復のことである。

この**反復**という用語は、ある条件下でデータを取る場合に (当然、異なる個体から) 複数回データを取る必要がある、ということを意味している。これは得られているデータが 1 つだけだと、真値と誤差の分離ができないためである。

フィッシャーの 3 原則のうち、反復については参考文献 [16] の解説が特に分かりやすい。反復を含めた他の用語については参考文献 [19] を参照するとよいだろう。ただし、参考文献 [19] では反復の説明のところで、ある条件下で複数回にわたってデータを取ることを「繰り返し」と述べているが、これは前述した繰返測定データの「繰り返し」とは意味が異なるので注意が必要である。

このような実験計画に関する参考書を読むと、読者は「たかがデータを取るために」と煩わしさを感じるかもしれない。しかし、データ解析において、得られたデータを解析するという作業は、コンピュータが発達している現代においてはそれほど大変な手間とはならないだろう。このような環境のなかでデータ解析が上手くいくかどうかは、解析の作業自体にあ

るのではなく、緻密な実験計画と厳密な測定作業の段階で左右されるといってもよい。それだけ解析作業の前段階は重要だということを強く主張したい (残念ながら心理学ではこの段階の重要性が軽視されていることが多い)。

- 独立なデータ (対応なしのデータ)　＝　反復測定データ
- 対応ありのデータ　＝　繰返測定データ or 経時測定データ

第 3 章

一般線形モデル

3.1　データフレーム

一般線形モデル、すなわちモデル解析を学ぶ前に**データフレーム**について知っておかなければならないだろう。t 検定で用いたデータセットである表 2.1 や分散分析で用いたデータセットである表 2.2 は、少なくともモデル解析の観点からすれば適切なデータ形式ではない。

例えば A 群、B 群、C 群の平均値を比較するために得られたデータセットについて考えてみよう。これは典型的な 1 要因 3 水準の分散分析のデータセットである。これは表 2.2 のように、1 列目に A 群のデータ、2 列目に B 群のデータ、そして 3 列目に C 群のデータとしてまとめる方法もありえる。一方でデータフレームで表現するということは表 3.1 のようにまとめるということである。

表 3.1(データフレーム) では 1 列に 1 つの変数が表現される形式になっている。そのためデータフレームでは、自分が解析に用いる変数が何であるかを明確にまとめられるという利点がある。またデータフレームでは 1 行に 1 人の被験体についてのデータがまとめられているので、サンプルサイズも明らかである (サンプルサイズは行数である)。さらにモデル解析で重要な連続型とカテゴリカル型の区別をつけるためにも役立つ。$Score$ は連続型、$Group$ はカテゴリカル型であることは一目瞭然であろう。

3.2　1 要因 2 水準の分散分析モデル

t 検定は一般線形モデルに含まれる**分散分析モデル**として解析することができる。ここで表 2.1 のデータセットを改めてデータフレームとしてまとめ直した表 3.2 を用いてモデル解析を行ってみよう。

モデル解析は自分が明らかにしようとする問題を**モデル式**で表すことから始める。ここで明らかにしたい問題とは「男性と女性とで身長の平均値が異なるか」ということである。これを換言すると「$Height$ の変動は Sex によって説明されるか」といえる。この問題をモデル式で表現すると、式 3.1 のように書くことができる。

表 3.1　データフレーム

Subject	Score	Group
No.1	56	A
No.2	60	A
No.3	55	A
No.4	43	B
No.5	70	B
No.6	50	B
No.9	33	C
No.10	20	C
No.11	15	C

表 3.2　男女の身長データ (データフレーム)

Subject	Height	Sex
No.1	175	Male
No.2	180	Male
No.3	169	Male
No.4	170	Male
No.5	173	Male
No.6	160	Female
No.7	155	Female
No.8	167	Female
No.9	170	Female
No.10	158	Female

$$Height = Sex \tag{3.1}$$

式 3.1 の意味を確認しておくと、これは「$Height$ は Sex によって説明される」ということを表している。$Height$ は応答変数で連続型、Sex は説明変数でカテゴリカル型である。説明変数が 1 つでカテゴリカル型なので、これは 1 **要因の分散分析モデル**である。

ところで分散分析において、カテゴリカル型の変数のことを特に**要因**とか**因子**と呼ぶということを覚えておくと良いだろう。ただし、ここでいう因子と因子分析でいうところの因子とは全く別の意味なので、混同しないように注意すべきである。

それでは実際に R を用いて式 3.1 を解析してみることにする。R では `lm()` という線形モデルを扱う関数を使うことになる (lm とは Linear Model の頭文字である)。R の優れている

点は式 3.1 のようなモデル式に忠実な表現方法でモデル指定を行えることである。

R での実行例は以下のボックスに示してある。しかし本書では R の実行例について詳しく解説することはしない。その理由は読者が実際に R を操作しながら「このコードはどういう意味なんだろう」と試行錯誤することによって R の上達がはやまると考えるからである。それゆえに、本書では R の解析例をあえて不親切な形で提示することにした。

例えば初心者にとっては `Sex <- rep(c(1, 2), c(5, 5))` というコードの意味さえ最初は分からないだろう。しかし変数 Sex の中身を確認してみると、`rep()` という関数がどのようなことをしてくれるものなのか、なんとなく分かるはずである。それから実行例の数値とは異なる数値を指定してみたりして、その場合にはどのような処理がなされるのかを確認することで理解できるはずだ (例えば `rep(c(1, 3), c(2, 4))` としてみたらどうなるだろうか)。

```
> Height <- c(175, 180, 169, 170, 173, 160, 155, 167, 170, 158)
> Sex <- rep(c(1, 2), c(5, 5))
> Sex
 [1] 1 1 1 1 1 2 2 2 2 2
> Sex <- factor(Sex, levels=c(1, 2), labels=c("Male", "Female"))
> Sex
 [1] Male  Male  Male  Male  Male  Female Female Female Female Female
Levels: Male Female
> result <- lm(Height ~ Sex)
> summary(result)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  173.400      2.425  71.509 1.63e-12 ***
SexFemale    -11.400      3.429  -3.324   0.0105 *

> summary(aov(result))
            Df Sum Sq Mean Sq F value  Pr(>F)
Sex          1  324.9   324.9  11.051 0.01047 *
Residuals    8  235.2    29.4
```

`summary(result)` によって**係数表**が得られる。この結果より、式 3.1 に定数項と Sex の係数を書き加えると式 3.2 のようになる。

$$Height = 173.4 + \left\{ \begin{array}{c} 0 \times Male \\ -11.4 \times Female \end{array} \right\} \tag{3.2}$$

式 3.2 の見方は簡単である。$Male$ の係数は 0 になっているが、これは $Male$ の平均値を基準としているためである。それに対して $Female$ の係数は -11.4 である。これは $Male$ より -11.4 だけ平均値が小さいことを意味している。

定数項 (Intercept) は 173.4 となっているが、実はこれが $Male$ の平均値である。すると $Female$ の平均値は $Male$ よりも -11.4 だけ小さいのだから、$173.4 - 11.4 = 162.0$ となるわけである。

モデル解析によって係数表が得られることのメリットとして、ある水準を基準として別のある水準の平均値を直接的に比較することができるという点があげられる (今回の例では男性という水準を基準 0 とした)。

さて一方で `summary(aov(result))` によって**分散分析表**が得られる。これは変数 Sex が変数 $Height$ の変動をどれだけよく説明できたかを示したものである。直感的にいってしまうと `Sum Sq.`(SS, Sum of Squares) は Sex が $Height$ をどれだけよく説明できたかを示しているものである。

つまり SS(平方和) が大きいということは、それだけ応答変数をうまく説明できたということである。それに対して `Residuals` はうまく説明できなかった部分である。この説明から Sex の SS はより大きく、Residulas の SS はより小さい方がよいということが直感的に分かるだろう。

なお `Mean Sq` というのは平均平方 (MS, Mean of Squares) と呼ばれるものであり、Sex の平均平方を Residulas(残差) の平均平方で割ったものが `F value`(F 値) となる ($324.9/29.4 = 11.051$)。そして p 値は $p = 0.01047$ となっていることから、Sex は有意に $Height$ の変動を説明していると主張することができる。

以上の分散分析に関する詳しい解説は 6 章で述べているのでそちらを参照してもらいたい。

3.3　1 要因 3 水準の分散分析モデル

前節に続いて、ここでも 1 要因の分散分析モデルを扱う。前節は古典的に t 検定として扱われる場合のデータセットを用いた。今度は分散分析として扱われる 3 群の平均値の差の検定についてモデル解析を適用してみる。

表 3.3 は 2.2 をデータフレームとしてまとめ直したものである。ここにある変数は $Score$(総合得点) と $Class$(組) の 2 つである。$Score$ は連続型、$Class$ は 3 水準のカテゴリカル型である。

ここでの問題は「組によって総合得点の平均値が異なるか」ということである。これは「$Score$ は $Class$ によって説明されるか」ということである。この問題をより単純明瞭に表現したものが式 3.3 のようなモデル式である。

表 3.3　総合得点と組のデータフレーム

Subject	Score	Class
No.1	374	Matsu
No.2	325	Matsu
No.3	203	Matsu
No.4	329	Matsu
No.5	228	Matsu
No.6	354	Matsu
No.7	417	Matsu
No.8	461	Take
No.9	361	Take
No.10	316	Take
No.11	282	Take
No.12	423	Take
No.13	328	Take
No.14	397	Take
No.15	356	Ume
No.16	285	Ume
No.17	437	Ume
No.18	315	Ume
No.19	415	Ume
No.20	381	Ume
No.21	324	Ume

$$Score = Class \tag{3.3}$$

式 3.3 の解析例を以下のボックスに示す。

```
> Score <- c(372, 325, 203, 329, 228, 354, 417,
+ 461, 361, 316, 282, 423, 328, 397,
+ 356, 285, 437, 315, 415, 381, 324)
> Class <- rep(1:3, c(7, 7, 7))
> Calss <- factor(Class, levels=c(1, 2, 3),
+ labels=c("Matsu", "Take", "Ume"))
```

```
> anova.model <- lm(Score ~ Class)
> summary(anova.model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   318.29      24.90  12.783 1.81e-10 ***
ClassTake      48.57      35.21   1.379    0.185
ClassUme       40.71      35.21   1.156    0.263

> summary(aov(anova.model))
            Df Sum Sq Mean Sq F value Pr(>F)
Class        2   9517    4758  1.0965 0.3553
Residuals   18  78112    4340
```

まず分散分析表の方から見ていってみよう。今回の結果は前節の 1 要因 2 水準の分散分析モデルの結果とは対称的である。$Class$ の SS は 9517 であり、Residuals の SS は 78112 である。これは $Class$ が説明する量より、Residulas の説明量の方が大きいことを表している。したがって、$Class$ は $Score$ に対して説明力を持たないということがいえる ($p = 0.3553$)。

続いて係数表を見てみよう。前節では説明しなかったが、この係数表における p 値はいったい何について検定したことによって得られるものであろうか。これは「推定された係数は 0 である」という帰無仮説について検定しているのである。定数項 (`Intercept`) については「推定された定数項は 0 である」という帰無仮説について検定していることになる。

`ClassTake`(竹組) の p 値は $p = 0.185$ であるから、例えば 5% の有意水準では帰無仮説を棄却して係数は 0 であることを意味している。同様に `ClassUme`(梅組) も $p = 0.263$ なので係数は 0 であるといえる。言い換えれば、これは推定された係数 (Take: 48.57, Ume: 40.71) は平均値の推定に役立たないということである。

ここでモデル式に係数を当てはめた式 3.4 を見てみよう。これを見れば、全ての水準についての係数が 0 であることが分かる。$Matsu$ を基準としているわけであるから、$Take$ と Ume の平均値は基準 0 と変わらないことを意味している。つまり 3 つの水準間で平均値に差がないということである。

$$Score = 318.29 + \left\{ \begin{array}{c} 0 \times Matsu \\ 0 \times Take \\ 0 \times Ume \end{array} \right\} \tag{3.4}$$

3.4　2 要因の分散分析モデル

3.4.1　加法モデルと交互作用モデル

分散分析モデルにおいて、2 つ以上の説明変数 (要因) を用いる場合には交互作用効果について考慮しなければならないだろう。2 **要因の分散分析モデル**では、まず交互作用項が有意であるかどうかを確認しなければならない。なぜならば、交互作用が認められるかどうかでその後の解釈の仕方も、論文に載せるべき適切なグラフ表現も変わってくるからである。

本書では交互作用項が含まれる場合の分散分析モデルを特に**交互作用モデル**と表現している。また一方で、交互作用項が含まれないモデルを**加法モデル**と区別して扱っている。

応答変数 Y を説明する変数として要因 A と要因 B があるとしたら、交互作用モデルは式 3.5 のように表される。

$$Y = A + B + A : B \tag{3.5}$$

一方で交互作用が有意でない、加法モデルは式 3.6 のように表現されることになる。

$$Y = A + B \tag{3.6}$$

3.4.2　加法モデルの解析と結果の解釈

まずは交互作用項が有意でない場合の例をあげて説明する。表 3.4 は 1 日の食事回数 (Num) と 1 日の中で間食をするかどうか ($Snack$) によって BMI(Body Mass Index) がどのように変化するかをまとめたものである。

分析の手順としては、(1) まず交互作用図を作成し、(2) フルモデルを解析して交互作用項が有意であるかを調べ、(3) 交互作用項が有意ならば交互作用図を、有意でなければ主効果についてのグラフを作成するという段階をふむことになる。

交互作用図

交互作用図を確認することによって、交互作用が認められるかどうかを視覚的に判断することができる。図 3.1 は横軸に 1 日の食事回数 ($Num : Levels = Zwei, Drei$) を、縦軸に BMI をとってある。そして間食の有無 ($Snack : Levels = No, Yes$) の水準ごとに 2 つの折れ線として表現してある。白いデータポイントが間食なし、黒いデータポイントが間食ありである。

交互作用図において、交互作用が認められる場合は 2 本の折れ線が有意に異なる (視覚的に平行ではなくなる)。それに対して交互作用がない場合は 2 本の折れ線が、視覚的にほぼ平行になる。ただし、交互作用図による判断はあくまでも目見当であるから、本当に有意であるかどうかは分散分析によって明らかにしなければならない。

表 3.4　食事回数と間食による BMI の違い

食事回数 (Num)	間食の有無 ($Snack$)	BMI
Zwei	No	19.5
Zwei	No	23.3
Zwei	No	26.0
Zwei	Yes	29.5
Zwei	Yes	28.9
Zwei	Yes	30.1
Drei	No	22.3
Drei	No	20.5
Drei	No	21.2
Drei	Yes	26.1
Drei	Yes	24.0
Drei	Yes	27.7

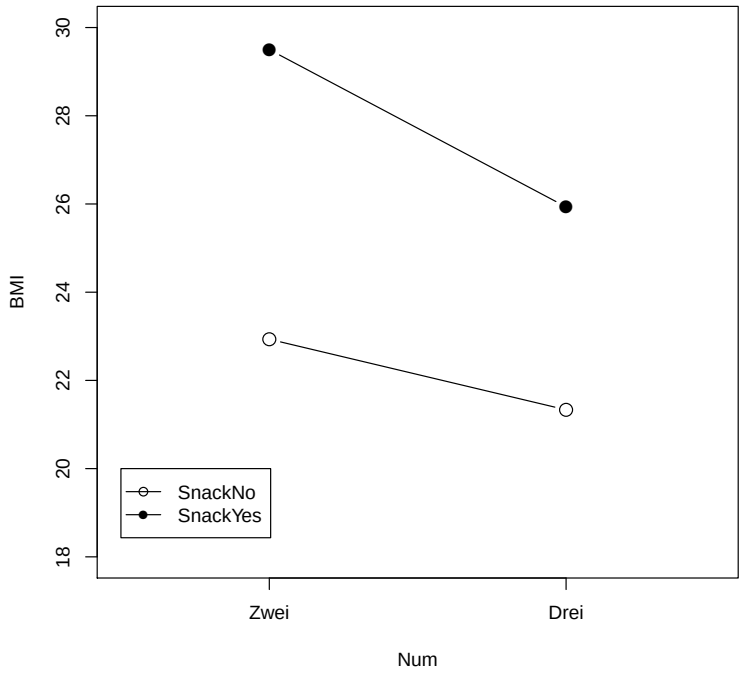


図 3.1　交互作用がない場合の交互作用図

ちなみに図 3.1 を見る限りでは、交互作用が認められないであろうことが分かる。2 本の折れ線は完全に並行ではないが、極端に異なっているわけでもないからである。

フルモデルからの解析

分散分析は最初に交互作用について考えなければならない。なぜならば、交互作用効果には実質的に主効果も含まれているからである。つまり、ある交互作用が有意であるならば、それに含まれる主効果が有意であろうとなかろうと重要な要因だということである。

このことからも、モデル解析については最も複雑なモデル (フルモデル) から解析をすることが適切であるといえるだろう。表 3.4 のデータセットを解析すると以下のような結果が得られる。

$BMI = Num + Snack + Num : Snack$ の解析
($Num, Snack$ はカテゴリカル型)

```
> tw.anova <- lm(BMI ~ Num + Snack + Num:Snack)
> summary(aov(tw.anova))
            Df Sum Sq Mean Sq F value   Pr(>F)
Num          1 20.021  20.021  5.2376 0.051378 .
Snack        1 93.521  93.521 24.4659 0.001126 **
Num:Snack    1  2.901   2.901  0.7589 0.409058
Residuals    8 30.580   3.822
```

これによれば `Num:Snack` という交互作用項が有意ではないことが分かる ($p = 0.409058$)。主効果についてはどうだろうか、もし有意水準を 5% としているなら `Num` は有意ではない ($p = 0.051378$)。しかし `Snack` は有意である ($p = 0.001126$)。

このことから交互作用項はモデルに不要な項であることが分かる。したがって交互作用項を除いたモデルに更新する必要があるだろう。R には新たなモデルへと更新するために `update()` という便利な関数が用意されている。

$BMI = Num + Snack$ の解析
($Num, Snack$ はカテゴリカル型)

`update()` **の使い方**:

```
update(model, ~. term)
```

`model` **には元のモデルを指定する。**

`~.` **は指定したモデルの全ての項を表している。**

term には削除・追加する項を指定する（-なら削除、+ なら追加）

```
> #tw.anova から Num:Snack を削除して更新する
> tw.anova2 <- update(tw.anova, ~. -Num:Snack)
> summary(aov(tw.anova2))
            Df Sum Sq Mean Sq F value   Pr(>F)
Num          1 20.021  20.021  5.3818 0.045495 *
Snack        1 93.521  93.521 25.1394 0.000725 ***
Residuals    9 33.481   3.720
```

すると今度は Num は有意になり ($p = 0.045495$)、$Snack$ の p 値もより小さな値となった ($p = 0.000725$)。この結果は $Num : Snack$ という交互作用項が他の項である Num や $Snack$ の説明量を減らしていたということを示している。ただし、p 値が有意水準より大きいか小さいかだけである変数を削除するかどうかを安易に判断してはならない。

有意水準を $\alpha = 0.05$ として検定を行った場合、仮にある変数についての p 値が $p = 0.07$ といったように、有意水準よりもわずかに大きかったとする。このとき分析者はこの変数を削除するべきか否かを迷うかもしれない。あるいは何の迷いもなく削除してしまうかもしれない。しかし、もし予測 (説明) しようとしている応答変数に対して重要な変数であると考えるならば、たとえ有意でなくてもその変数を削除するべきではないこともある。

例えば医薬品の効果を調べるような場面では 0.02 の違いは無視できない重要な違いだろう。しかしマーケティングや世論調査のような場面においては、ある程度は融通を利かせたほうがよいかもしれない。わずかな p 値の違いで実質的に重要な変数を削除することの方が、もしかしたら危険かもしれない。データ解析は単なる数値比較をすることではないので、常に実質科学的な考えをもって臨まなければならないのである。

加法モデルの結果の表示

今回の例では交互作用項が有意ではなく、加法モデルであることが示された。交互作用が認められない場合、結果はそれぞれの主効果のみを示せばよい。つまり Num による平均値の違いと $Snack$ による平均値の違いは別々のグラフで表現するのである (図 3.2、図 3.3)。

モデル式への係数の当てはめ

最後にモデル式に推定した係数を当てはめて整理するとよい。本書ではモデル式とモデル式に係数を当てはめた式を区別するために、これを**係数式**と呼ぶことにする。

今回の解析より係数式は式 3.7 のように書けばよいことになる。1 要因の分散分析より係数式も複雑になるが、係数式の見方は同じである。定数項である 23.42 という値は $NumZwei$

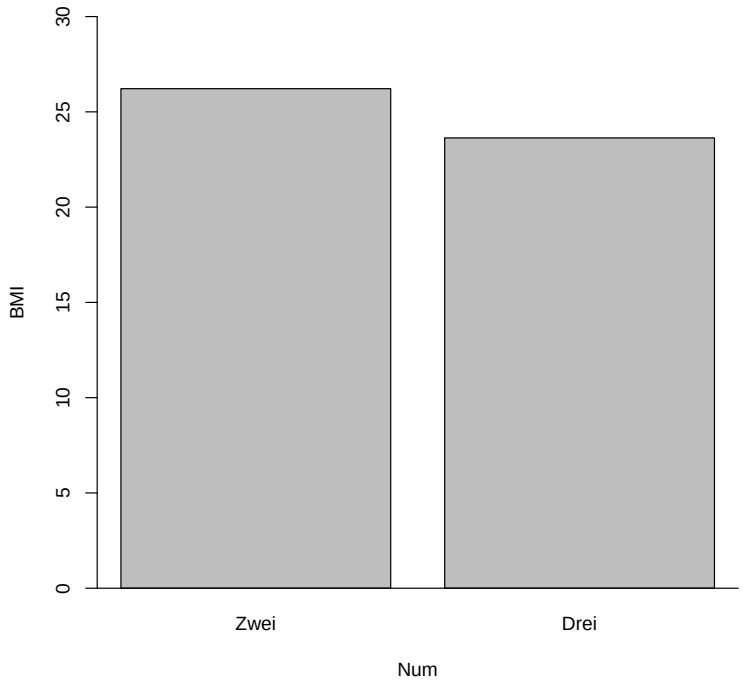


図 3.2　Num の主効果

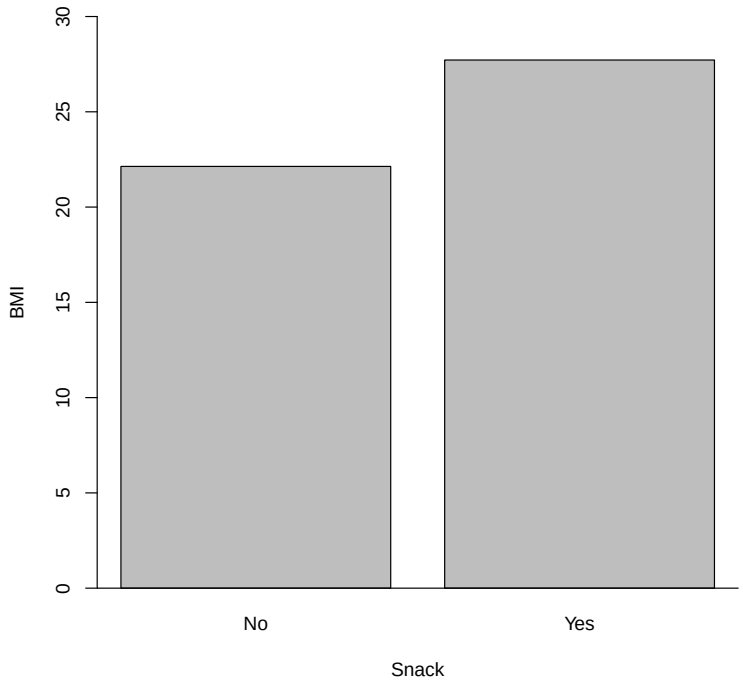


図 3.3　$Snack$ の主効果

かつ $SnackNo$ の平均値である。同様に考えて、例えば $NumDre$ かつ $SnackNo$ の平均値は $23.42 - 2.58 + 0 = 20.84$ と計算できる。

$$BMI = 23.42 + \left\{ \begin{array}{c} 0 \times NumZwei \\ -2.58 \times NumDrei \end{array} \right\} + \left\{ \begin{array}{c} 0 \times SnackNo \\ 5.58 \times SnackYes \end{array} \right\} \tag{3.7}$$

3.4.3　交互作用モデル

ここで新たに表 3.5 のデータセットを用いて 2 要因の分散分析モデルの解析例を示す。このデータセットはある洋食店の照明色 ($Light$) に対しての印象 ($Impression$) を男女別 (Sex) に 10 点満点で評価してもらったデータをまとめたものである。

表 3.5　照明の色に対する評価

Sex	Light	Impression
male	orange	6
male	orange	5
male	orange	6
male	blue	4
male	blue	4
male	blue	5
female	orange	9
female	orange	8
female	orange	8
female	blue	4
female	blue	5
female	blue	2

解析に先立って交互作用図を描いてみると図 3.4 のようになる。実はこの交互作用図は R の `intaraction.plot()` という関数を用いると簡単に描くことができる。以下にデータフレームの作成法と共にその手順を示す。

```
> x1 <- rep(1:2, c(6, 6))
> x1 <- factor(x1, levels=c(1, 2), labels=c("male", "female"))
> x2 <- rep(c(1, 2, 1, 2), c(3, 3, 3, 3))
> x2 <- factor(x2, levels=c(1, 2), labels=c("orange", "blue"))
> y <- c(6, 5, 6, 4, 4, 5, 9, 8, 8, 4, 5, 2)
```

```
> # データフレームの作成
> mydata <- data.frame(Sex = x1, Light = x2, Impr = y)
> mydata
      Sex  Light Impr
1    male orange    6
2    male orange    5
3    male orange    6
4    male   blue    4
5    male   blue    4
（以下省略）

> # パスを通す
> attach(mydata)

> # 交互作用図を描く
> interaction.plot(Sex, Light, Impr, type="b", pch=c(12, 16))

> # パスを解除
> detach(mydata)
```

さて Sex と $Light$ の交互作用図 (図 3.4) を見てみると、$female$ かつ $orange$ という水準の組合せの元で $Impr$ の平均値が極端に高くなっていることが分かる。これは交互作用が認められる場合の典型例である。実際に解析してそれを確かめてみよう。

$Impr = Sex + Light + Sex : Light$ の解析
($Sex, Light$ はカテゴリカル型)

```
> LC.model <- lm(Impr ~ Sex + Light + Sex:Light)
> summary(aov(LC.model))
            Df  Sum Sq Mean Sq F value    Pr(>F)
Sex          1  3.0000  3.0000     3.6 0.0943498 .
Light        1 27.0000 27.0000    32.4 0.0004585 ***
Sex:Light    1  8.3333  8.3333    10.0 0.0133491 *
Residuals    8  6.6667  0.8333
```

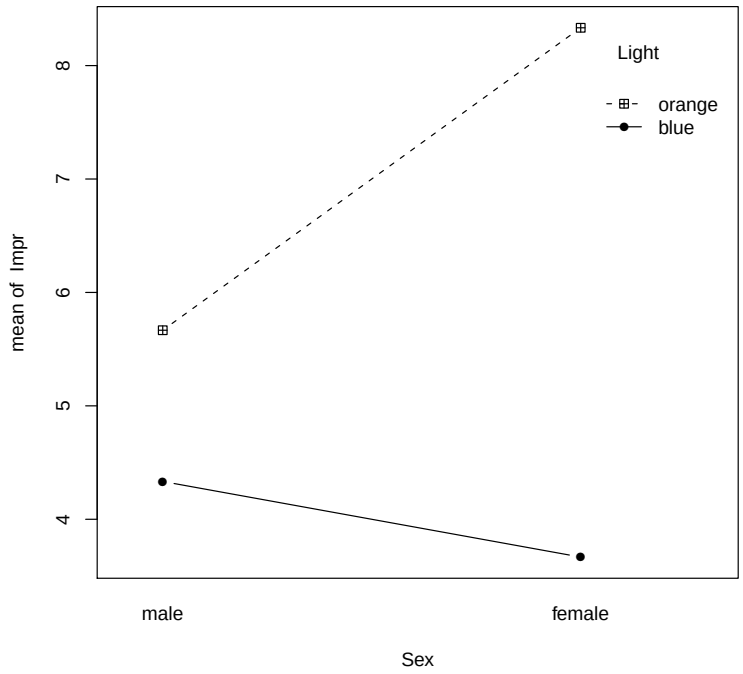


図 3.4　Sex と $Light$ の交互作用図

案の定、`Sex:Light` は有意である ($p = 0.0133$)。しかし `Light` も有意であるが ($p = 0.0004$)、`Sex` は有意ではない ($p = 0.0943$)。これはどのように解釈すればよいだろうか。

前節の加法モデルの説明を読んだならば、ここで有意ではない Sex を削除したモデルを解析しようとするかもしれない。しかしそれは間違いである。たとえ主効果が有意でなくとも、交互作用が有意であれば交互作用項に含まれる 2 つの要因は重要である。実質的に交互作用には 2 つの主効果が含まれているのである。

これは Sex が単独では効果がないが、$Light$ と組み合わさると効果があるということを示している。換言すれば Sex の効果は $Light$ の 2 つの水準のどれであるかに依存するといえる。この場合だと $female$ は $orange$ という水準であることに依存して、他の水準よりも高い値をとっていることになる。もし Sex が $Light$ のいずれの水準にも依存しないならば、図中の 2 本の折れ線は平行になるだろう (すなわち、それは加法モデルであることを意味している)。

交互作用モデルの結果の表示

さて、交互作用が認められる場合には Sex や $Light$ それぞれの主効果を別々に示すことには意味がなくなる。それゆえに、結果には交互作用図を載せればよいといえる。前述したように、Sex は $Light$ の水準に依存しているので、Sex のみの効果を主張したところで意味はないのである。

係数式には交互作用を含めた全ての項に推定値 (係数) を当てはめて書くことになる (式

3.8)。交互作用項を含めた係数式は複雑なものとなる。

$$
\begin{aligned}
Impression = 5.66 &+ \left\{ \begin{array}{c} 0 \times Sexmale \\ 2.66 \times Sexfemale \end{array} \right\} \\
&+ \left\{ \begin{array}{c} 0 \times Lightorange \\ -1.33 \times Lightblue \end{array} \right\} \\
&+ \left\{ \begin{array}{c} 0 \times male : orange \\ 0 \times female : orange \\ 0 \times male : blue \\ -3.33 \times female : blue \end{array} \right\}
\end{aligned}
\tag{3.8}
$$

交互作用の解釈における問題

先ほど「$female$ は $orange$ という水準であることに依存して、他の水準よりも高い値をとっている」という説明をしたが、この解釈は正しいだろうか。式 3.8 の交互作用項の部分を見てみると、$female : blue$ の係数のみが -3.33 となっている。これは「$female$ は $blue$ という水準の下で低い値をとる」ということができる。これらはどちらも同じことをいっているが、分析者がどのように解釈するか (表現するか) によって論文などの読者に与える印象は違ってくるだろう (もっとも、この出力例は自由に変えることができる。これについては 3 **章** 5 **節 ダミー変数**を参照せよ)。

オレンジ色の照明が高い評価を得るのか、それとも青色の照明が低い評価を得るのか、という解釈問題である。現実場面に応用する場合、前者が正しいならば店内の照明色をオレンジにすれば高い集客効果が期待されるだろう。しかし、後者が正しいならばオレンジ色の照明を設置したところで集客効果があるかどうかは不明である。なぜならば、そもそもオレンジ色に対する評価は標準的なものであって、好印象を与えているとは限らないからである。

実はこのような問題は、基準となる照明色 (一般的な白色の蛍光灯) という水準も $Light$ に含めれば解決する。もしオレンジ色の照明が集客に効果的ならば、標準色の折れ線は、オレンジの折れ線よりも下に描かれるだろう (図 3.5 の下側)。逆にオレンジ色の照明が効果的でないならば、標準色とオレンジ色の折れ線はほぼ同じ位置に描かれるだろう (図 3.5 の上側)。この場合、オレンジ色の照明が集客に効果的であると考えるよりは、青色の照明が集客には役立たないと考えた方が妥当である。

このように、データ解析とは単にコンピュータで数値計算をする作業には限らない。以上に述べたようなことを実験計画の段階で考慮すべきである。もし実験計画の段階でそれを見落としていたとしても、解析結果よりそれが明らかになれば、次の実験への課題として対策法も明らかになるだろう。

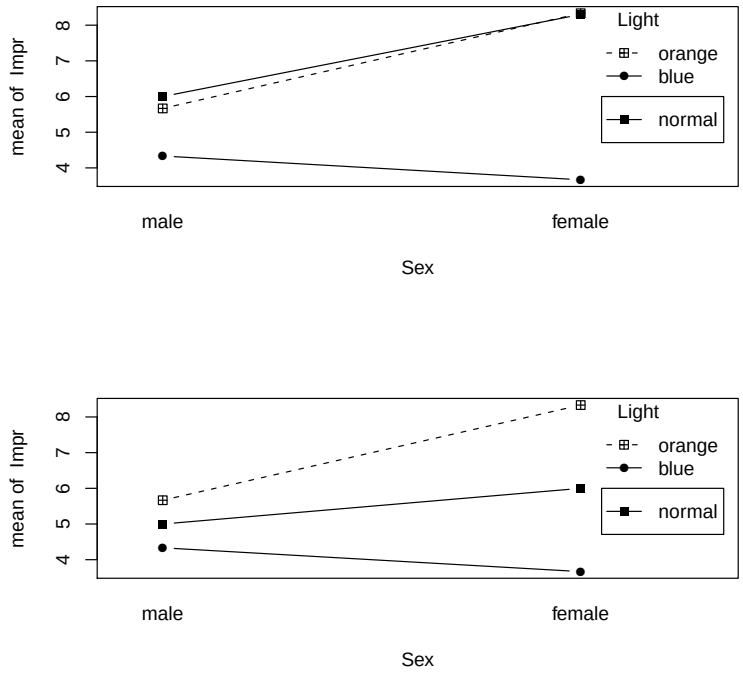


図 3.5　交互作用の解釈問題

3.5　ダミー変数

3.5.1　カテゴリカル型を連続型で表現する

ここまで分散分析モデルの解析例をいくつか見てきた。分散分析モデルは説明変数がカテゴリカル型であるが、R をはじめとする他の統計ソフトでは直接的にカテゴリカル型のデータを扱うことはできない。それではなぜ分散分析モデルが解析できるのだろうか。実はカテゴリカル型のデータは連続型のデータに変換されて解析が行われているのである。

まずは**ダミー変数**についての説明をすることにしよう。例えば性別という変数には男性と女性という 2 つの水準がある。これは数字の 0 と 1 を上手く組み合わせることによって、連続型のデータとして表現することができる。

性別のような 2 つの水準をもつカテゴリカル型のデータは特に 2 **値データ**と呼ばれている。そしてこの 2 値データは連続型 (間隔尺度) として扱ってよいことになっている。したがって、あらゆるカテゴリカル型データは 2 値データ (0 と 1) で表現できれば、それは連続型データとして解析することができるのである。

それでは実際に性別という変数 (カテゴリカル型) をダミー変数 (連続型) で表現してみよう。性別という変数を表でまとめると 1 列で表される (表 3.6 の第 2 列)。ダミー変数で表現すると、これは男性と女性という 2 列で表現されることになる (表 3.6 の第 3 列と第 4 列)。

表 3.6　性別のダミー変数による表現

被験者	性別	男性	女性
No.1	男性	1	0
No.2	男性	1	0
No.3	男性	1	0
No.4	女性	0	1
No.5	女性	0	1
No.6	女性	0	1

それではもう少し詳しく解説していこう。まず表 3.6 の第 3 列目にある男性の列を見てほしい。被験者 No.1 から No.3 は男性であるので「1」となっている。次に第 4 列目にある女性の列を見てみよう。当然だが男性であるということは、女性ではありえないので No.1 から No.3 は「0」となっている。このように考えると No.4 から No.6 の被験者は女性であることが分かる。男性の列には「0」、女性の列には「1」と入力されているからである。

実際に解析に用いるときには、男性か女性どちらか 1 つのダミー変数のみを使えばよい。なぜなら、どちらか一方だけでも被験者が男性であるか女性であるかを判断できるからである。例えば、男性の列は 1, 1, 1, 0, 0, 0 となっているが、「1」に該当すれば男性であることが分かり、水準は 2 つしかないので「0」である場合は必然的に女性であることが確定するのである。

ここで改めて表 3.2 のデータセットを用いて、今までどおりカテゴリカル型として扱う場合とダミー変数を用いて解析した結果を比較してみよう。

```
> Heigh <- c(175, 180, 169, 170, 173, 160, 155, 167, 170, 158)

> # カテゴリカル型の変数を用意する
> Sex <- rep(c(1, 2), c(5, 5))
> Sex <- factor(Sex, levels=c(1, 2), labels=c("Male", "Female"))

> # 説明変数はカテゴリカル型
> result <- lm(Height ~ Sex)
> summary(result)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  173.400      2.425  71.509 1.63e-12 ***
SexFemale    -11.400      3.429  -3.324   0.0105 *

--- --- ---

> # Male と Female というダミー変数を用意する
> Male <- rep(c(1, 0), c(5, 5))
> Male
 [1] 1 1 1 1 1 0 0 0 0 0
> Female <- rep(c(0, 1), c(5, 5))
 [1] 0 0 0 0 0 1 1 1 1 1

> #説明変数は連続型（ダミー変数 Female を使用）
> result2 <- lm(Height ~ Female)
> summary(result2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  173.400      2.425  71.509 1.63e-12 ***
Female       -11.400      3.429  -3.324   0.0105 *
```

以上の結果より、Sex を説明変数に指定した場合と Female を指定した場合とで、全く同じ結果が得られていることが分かるだろう。上の例では男性の列 (Male) を削除して解析を行ったが、これは男性の平均値を基準 (0) としたということである。つまり定数項が男性の平均値で (173.4)、女性の平均値は定数項から Female の係数を引いたものである ($173.4 - 11.4 = 162$)。これは 1 要因 2 水準の分散分析モデルのところで解説したことと同じであることを思い出してもらえただろうか。

ところで性別は 2 水準だが、3 水準以上の場合はどうなるだろうか。これも原理は同じで、例えば学校区分 (小学、中学、高校) という 3 水準のカテゴリカル型変数を連続型で表現してみると、それは表 3.7 のように表現される。これについても、小学、中学、高校いずれかの列を削除しても被験者がどの学校区分であるかを判断できるので、解析の際にはどれか 1 つを削除して行うことになる。

表 3.7 学校区分のダミー変数による表現

被験者	性別	小学	中学	高校
No.1	小学	1	0	0
No.2	小学	1	0	0
No.3	中学	0	1	0
No.4	中学	0	1	0
No.5	高校	0	0	1
No.6	高校	0	0	1

3.5.2 R における対比行列

R のデフォルト

上であげた身長と性別の解析例で気づいたかもしれないが、R の `lm()` でカテゴリカル型の変数を指定するとデフォルトでは第 1 列 (第 1 水準) が削除されて解析される。削除されるとは、その水準を**ベースライン**として解析されるということである。

少し専門的な表現をすれば、ある変数が k 個の水準を持っている場合、$k-1$ の線形結合で置き換えられるという。そしてこのような線形結合の選択法を**対比**という。つまり R のデフォルトでは「最初の水準を削除する」という対比を用いているわけである。R には `poly, helmert, sum, treatment` という 4 つの対比が用意されてるのだが、デフォルトでは `treatment` が指定されているというわけである (それぞれの対比については参考文献 [12] を参照せよ)。

SAS と同じ出力を得る場合

R では SAS と同じ出力を得ることができる。以下の解析例を見れば明らかであるが、SAS では R とは逆に最後の水準を削除した対比を採用しているようである。

```
> like.SAS <- lm(Height ~ Sex, contrasts=list(Sex="contr.SAS"))
> summary(like.SAS)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  162.000      2.425  66.808 2.80e-12 ***
SexMale       11.400      3.429   3.324   0.0105 *
```

```
--- --- ---

> summary(lm(Height ~ Male))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  162.000      2.425  66.808 2.80e-12 ***
Male          11.400      3.429   3.324   0.0105 *
```

Minitab と同じ出力を得る場合

Minitab では零和行列と呼ばれる対比を採用しており、これはダミー変数の係数パラメータ (推定値) に「和が 0」という制約をおいたものである。

```
> like.Minitab <- lm(Height ~ Sex, contrasts=list(Sex="contr.sum"))
> summary(like.Minitab)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  167.700      1.715  97.805 1.33e-13 ***
Sex1           5.700      1.715   3.324   0.0105 *
```

この場合の定数項は男性と女性を混みにした全平均となっている。そして Sex の係数である 5.700 という値は、全平均からどれだけ離れているかを表している。つまり男性の平均値は $167.7 + 5.7 = 173.4$ と計算することができる。前述したように係数の和が 0 になるように制約をつけているので、女性 (Sex2) の係数 β_{female} は $5.7 + \beta_{female} = 0$ として求められる。結局 $Sex2 = -5.7$ となるので、女性の平均値は $167.7 - 5.7 = 162.0$ と計算できる。

以上のような計算は dummy.coef() という関数を用いれば、わざわざ自分で計算する必要はなくなる (練習のために何回かは自分で計算してみるとよい)。

```
> dummy.coef(like.Minitab)
Full coefficients are
```

```
(Intercept):    167.7
Sex:             Male Female
                  5.7   -5.7
```

3.6 単回帰分析モデル

ここまで説明変数として用いられた変数はカテゴリカル型のみであった。ここからは連続型の変数を説明変数として扱う**回帰分析モデル**を紹介していく。説明変数として 1 つの連続型がおかれる場合、特に**単回帰モデル**と呼ばれるが、まずはこの単回帰モデルについて説明する。

単回帰モデルの最も典型的な例は「身長の変動は体重によって説明されるか」という問題について考えることである。この問題について、表 3.8 のデータセットを用いて解析手順を示す。

表 3.8 身長と体重のデータセット

Height	Weight
180	74
169	70
181	80
175	69
170	65
166	60
171	68
185	69
177	68
180	70
172	66

散々、繰り返してきたことであるがモデル解析において最初にすべきことは、分析者が取り扱う問題をモデル式で表すことである。今回の「身長の変動は体重によって説明されるか」という問題は式 3.9 のように書くことができる。

$$Height = Weight \tag{3.9}$$

取り扱う問題をモデル式で表現できたならば、次にやるべきことはデータを図示してみる

ことである。データを図示することによって、自分が解析しようとしているモデルを当てはめることが妥当であるかどうかを把握することができる。

単回帰モデルは線形モデルなので、直線を当てはめることになる (曲線を当てはめるのは非線形モデルである)。したがって、データを図示することによって直線を当てはめることが適切であるか (直線を当てはめることによって応答変数の変動を説明できるか) を把握しておく必要がある。あるいは、**はずれ値**がないかを把握するためにもデータを図示することは役立つ。

例えば図 3.6 は**相関関係**が見られない、いわゆる無相関であるような場合には、直線を当てはめることに意味はないだろう。また別の例として、図 3.7 のような場合に直線を当てはめることは明らかに無理があるだろう。

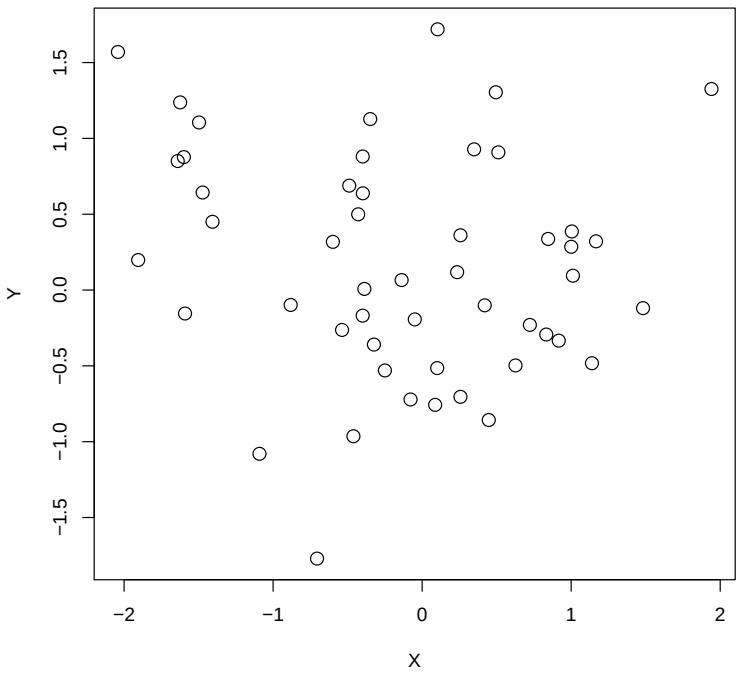


図 3.6　無相関である場合

線形モデルを扱う際に最も厄介な事態は図 3.7 のような曲線関係がみられる場合である。このような場合には対数変換などの変数変換を行うことによって、直線回帰を行うことが可能になることもある。これは扱う説明変数の数が多いときにより面倒な事態であるが、それについてはここでこれ以上は触れないこととする。

さて、少し話が逸れてしまったが、表 3.8 のデータを図示してみると図 3.8 のようになる。図を見ると曲線関係はなさそうなので、直線回帰を行っても問題ないだろう。

以下には単回帰モデル (式 3.9) の解析例を示す。

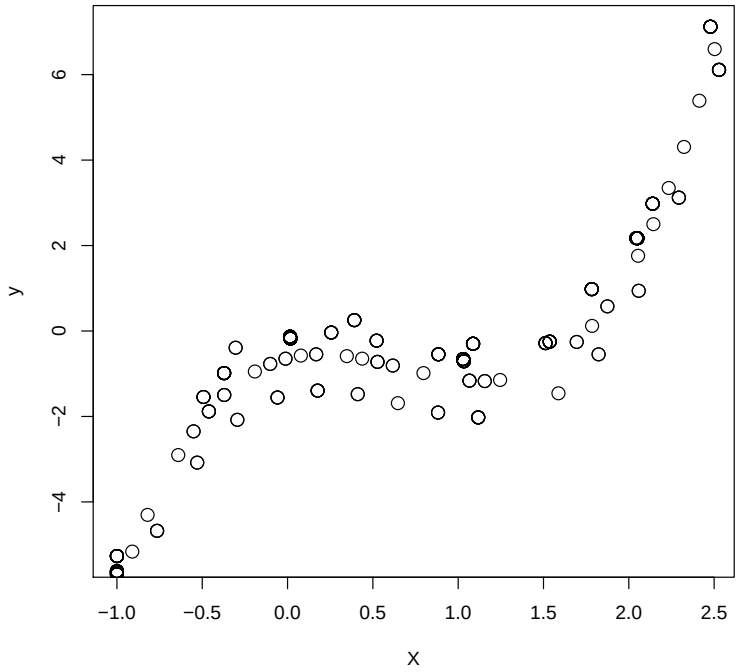


図 3.7　曲線関係がある場合

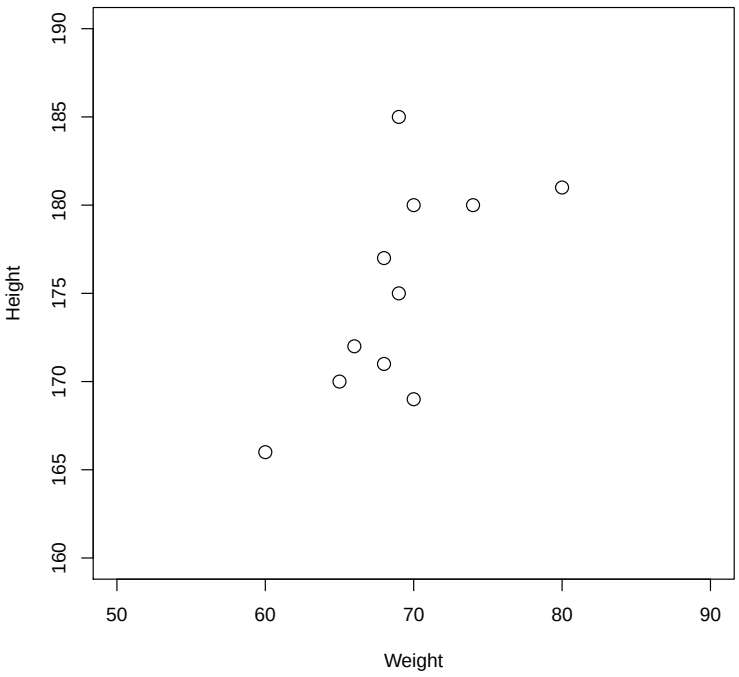


図 3.8　身長と体重の関係

$Height = Weight$
($Weight$ は連続型)

```
> result <- lm(Height ~ Weight)
> summary(aov(result))
            Df  Sum Sq Mean Sq F value  Pr(>F)
Weight       1 159.391 159.391  7.2627 0.02460 *
Residuals    9 197.518  21.946

> summary(result)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 120.6456    20.2521   5.957 0.000213 ***
Weight        0.7891     0.2928   2.695 0.024597 *
```

分散分析表を見てみると、Weight は有意であることがわかる ($p = 0.02460$)。回帰分析モデルの解析における分散分析表も分散分析モデルの解析で得られた分散分析表と同じことを意味するものである。復習としておさらいしておくと、Sum Sq は平方和、Mean Sq は平均平方であり、どちらも変動の具合を表している。SS(平方和) はデータ数によって値が変わるので、1 自由度あたりの変動を表す MS(平均平方) を F 値の計算に用いる ($F = MS_{Weight}/MS_{Residulas} = 159.391/21.946 = 7.2672$)。詳しくは 6 **章**の**回帰分析の原理**の章を参照してもらいたい。

続いて係数表 (Coefficients:) を見てみよう。定数項 (Intercept) と体重 (Weight) の係数が出力されている。Std.　Error は標準誤差であり、これは推定値 (Estimate) の推定精度を表す指標である (この値は 0 に近い値であることが望ましい)。これを用いて推定された値の**信頼区間**を求めることができる (6 **章**の**検定と推定**を参照せよ)。

また推定された値を標準誤差で割ることによって、検定統計量である t 値が得られる。この場合の帰無仮説は「定数項は 0 である」というものと、「$Weight$ の係数は 0 である」というものである。どちらも有意であるので、推定された値は予測に役立つということができる (p-value of Intercept $= 0.000$, p-value of Weight $= 0.024$)。

さて、得られた定数項と係数を式 3.9 に書き加えたものが式 3.10 である。これは**回帰方程式**と呼ばれているもので、応答変数 ($Height$) の予測値を求めることができるのである。

$$Height = 120.64 + 0.78 \times Weight \tag{3.10}$$

ちなみに予測値は predict() という関数を用いれば簡単に得ることができる。

```
> predict(result)
        1         2         3         4         5         6
 179.0362 175.8800 183.7706 175.0909 171.9347 167.9893
        7         8         9        10        11
 174.3018 175.0909 174.3018 175.8800 172.7237
```

また、予測値と観測値のズレを**残差**という。当然だが、残差の値が 0 であるということは、予測値と観測値のズレが全くないということであり、これは式 3.10 によって観測値を完全に予測できていることを意味する。換言すれば「説明変数 $Weight$ は応答変数 $Height$ を完全に説明できている」ということである。

回帰分析において、この残差という値について考えることは重要で、これは `residuals()` という関数を用いれば求めることができる。回帰分析が妥当であったかどうかを診断することを**回帰診断**というが、これについては重回帰モデルの節で詳しく解説する。

```
> residuals(result)
          1           2           3           4           5           6
 0.9637784 -6.8799716 -2.7705966 -0.0909091 -1.9346591 -1.9893466
          7           8           9          10          11
 -3.3018466 9.9090909  2.6981534  4.1200284 -0.7237216
```

3.7　重回帰分析モデル

重回帰分析モデルとは連続型の説明変数が 2 つ以上ある場合である。重回帰モデルを扱うにあたって、説明変数として用いる変数が多くなる場合に**モデル選択**（あるいは**変数選択**と言い換えてもよい）が重要となってくる。例えば 4 つ以上の説明変数がある場合には、全ての交互作用項を含めたモデルは非常に複雑なものとなってしまう。これは推定するパラメータが増加するということであり、そのためには大きなサンプルサイズ（データ数）が要されることになる。

こういった推定するパラメータの数が多くなりすぎるという問題もあるので、重回帰分析においては、最初から交互作用項を含めない加法モデルとして解析することがある。ただし、R には `step()` というモデル選択（変数選択）を自動的に行ってくれる関数が用意されている。これを用いれば交互作用項を含めた複雑なモデルも（言い方は悪いが）勝手に適切なモデルを作り上げてくれる。しかし、`step()` によって作られたモデルが必ずしも最適なモデ

ルであるとは限らないことも気に留めておくべきであろう。

ここでは R に最初から実装されているデータセットである、swiss を用いて重回帰モデルの解析を紹介する。また、分析者自信が手作業でモデル選択を行う場合と、step() を利用して得られたモデルとを比較してみよう。

3.7.1 ユーザの手作業による解析例

R には最初からいくつかのデータセットが用意されており、その一覧を見るためには data() とコマンドすればよい。実際に一覧を見てみると swiss のデータについて「Swiss Fertility and Socioeconomic Indicators (1888) Data」という説明書きを見つけることができるだろう。これは 1888 年に取られたスイスにおける出生と社会経済指標についてのデータセットであり、これには次のような変数が含まれている（データセットを data(swiss) とコマンドして呼び出した後、names(swiss) とコマンドすれば変数名が表示される）。

- Fertility：　出生率
- Examination：　軍事テストで最高点を取った被徴兵者の割合
- Catholic：　カトリック教徒の割合
- Infant.Mortality：　1 歳未満の幼児死亡率
- Agricalture：　農業従事者
- Education：　被徴兵者の小学校以上の学歴

散布図の確認

いかならデータ解析においても、まず最初にデータをプロットしてみることが重要である。特に重回帰分析においては扱う変数が多くなることがあるので、使用する変数間にどのような関係が見られるかを視覚的に判断することが第一の作業となる。明らかに応答変数と関係のなさそうな変数はモデルに含める必要はないし、大きなはずれ値が存在しているような場合にはそれを発見することができる。

R には pairs() という便利な関数が用意されている。これを用いることによって、使用する変数全ての組合せについて散布図が出力される（図 3.9）。また、引数に panel=panel.smooth と指定することでスムージング曲線が引かれる。これはデータの散布度を視覚的に捉えるのに便利である。

```
> data() # データセットの一覧を表示
> data(swiss) # swiss の呼び出し
> names(swiss) # 変数名を確認
[1] "Fertility"       "Agriculture"     "Examination"     "Education"
```

```
[5] "Catholic"        "Infant.Mortality"
> pairs(swiss, panel=panel.smooth) # 散布図行列の出力
```

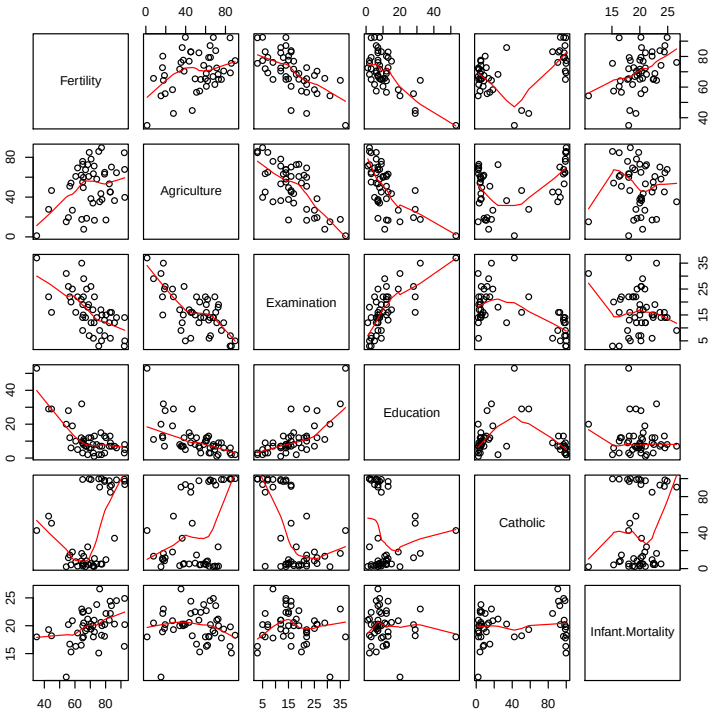


図 3.9　swiss データセットの散布図行列

曲線関係の探索

応答変数と説明変数との間に曲線関係があるかどうかを探索することは重要である。なぜなら、もし曲線関係が疑われるようならば、モデルに 2 次項 (自乗を含む項) を含める必要が出てくるからである。なお、2 次の交互作用項 $(X_1 : X_2)$ と 2 次項 (X_1^2) とは名前が似ているが違うので注意してほしい。

さて、曲線関係を探索するためには `gam()` という関数を用いるのが便利である。これは一般化加法モデル（Generalized Additive Model）のための関数であるが、この手法についての詳しい仮説は割愛する（詳しくは文献 [12] を参照のこと）。それで、この `gam()` という関数を使うには `library()` という関数を用いてライブラリ（パッケージが置いてある倉庫のようなもの）から呼び出してやる必要がある。この際、後のロジスティック回帰分析で用いるパッケージ VGAM もインストールしておくとよい。

```
> install.packages("VGAM") # VGAM のインストール
> library(mgcv) # mgcv の呼び出し
> attach(swiss) # swiss にパスを通しておく
> # gam.result に gam() による解析結果を代入する
> gam.result <- gam(Fertility ~ s(Agriculture) +
+ s(Examination) + s(Education) +
+ s(Catholic) + s(Infant.Mortality))
> windows() # 新しいグラフウィンドウを起動
> par(mfrow=c(2, 3)) # ウィンドウを 2 行 3 列に分割
> plot(gam.result) # プロット
```

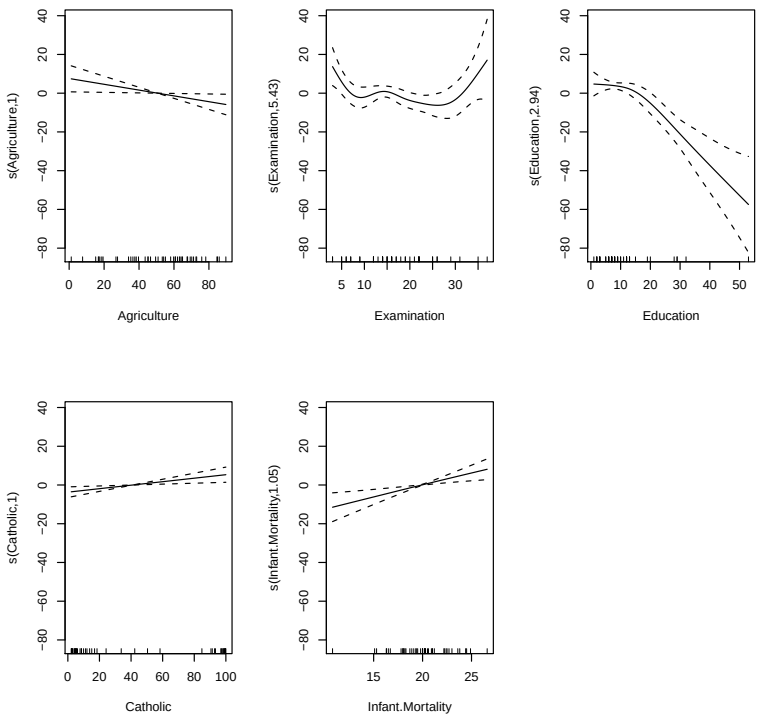


図 3.10　平滑化スプライン曲線

図 3.10 を見てみると、*Examination* と *Education* とに曲線関係があるようである。したがって、モデルに `Examination^2` と `Education^2` という 2 次項を組み入れて考える必要がありそうだということが分かる。

モデル選択

最初に曲線関係が疑われるので、交互作用項を含まないモデルに Examination^2 と Education^2 の 2 次項を入れて分析してみる。出力結果は係数表の部分のみで、あとの部分は省略してある。

```
> lm.res1 <- lm(Fertility ~ . + I(Education^2) + I(Examination^2))
> summary(lm.res1)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      65.59858   10.72032   6.119 3.54e-07 ***
Agriculture      -0.12237    0.07346  -1.666  0.10375
Examination      -1.52247    0.70188  -2.169  0.03623 *
Education        -0.37282    0.41586  -0.896  0.37549
Catholic          0.07795    0.03724   2.093  0.04289 *
Infant.Mortality  1.37971    0.40653   3.394  0.00159 **
I(Education^2)   -0.01241    0.00828  -1.498  0.14208
I(Examination^2)  0.03542    0.01854   1.910  0.06347 .
```

これによれば、どちらの 2 次項も有意ではないが Education^2 については $p = 0.14$、Examination^2 については $p = 0.06$ と、これらの項を取り除くには少々の不安が残る（gam() によって得られたグラフをみても、曲線関係がありそうなのは明らかである）。

そこで今度は 2 次項を含めたまま、交互作用項を含めたモデルを解析してみる。モデル式の記述で~ . という部分があるが、これは全ての変数を用いるという意味である。ピリオド（.）は全ての変数という意味である。また、それに^2 と付け加えられているのは 2 次の交互作用を指定するという意味で、2 次項（累乗項）を意味しているのではないことに注意すること。

```
> lm.result1 <- lm(Fertility ~ .^2 + I(Education^2) + I(Examination^2))
> summary(lm.result1)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                    259.679228  68.381854   3.797 0.000691 ***
Agriculture                     -1.966543   0.708890  -2.774 0.009581 **
Examination                     -8.943802   3.799544  -2.354 0.025565 *
Education                        0.568031   3.944942   0.144 0.886505
Catholic                        -0.380349   0.431479  -0.882 0.385294
Infant.Mortality                -5.904220   3.283151  -1.798 0.082541 .
I(Education^2)                  -0.039096   0.029672  -1.318 0.197956
I(Examination^2)                 0.010867   0.057384   0.189 0.851122
Agriculture:Examination          0.032384   0.017709   1.829 0.077747 .
Agriculture:Education            0.003455   0.019102   0.181 0.857714
Agriculture:Catholic             0.003454   0.002995   1.153 0.258170
Agriculture:Infant.Mortality     0.053439   0.031096   1.719 0.096365 .
Examination:Education            0.106770   0.063367   1.685 0.102733
Examination:Catholic            -0.004868   0.015071  -0.323 0.749038
Examination:Infant.Mortality     0.284293   0.160229   1.774 0.086515 .
Education:Catholic               0.001836   0.013028   0.141 0.888911
Education:Infant.Mortality      -0.123877   0.168556  -0.735 0.468286
Catholic:Infant.Mortality        0.015813   0.017787   0.889 0.381290
```

これ以降は各項の p 値を参考にしながら、複雑な交互作用項、あるいは 2 次項を優先的に削除していく（モデルの更新には `update()` を用いる）。なぜ複雑な項から削除していくのかというと、最終的に採用されるモデルはなるべく単純なモデルであるべきだからである。またこの際に、ある変数を削除した後にどの変数に対する p 値が増加（あるいは減少）したかを注意深く観察しなければならない。取り除いた変数は他の項の説明率を上げるのか下げるのかを見極めるということである（これは経験に依存するところが大きい）。

まず `Agriculture:Education`、`Examination:Catholic`、`Education:Catholic` は明らかに有意ではない（それぞれ $p = 0.857, p = 0.749, p = 0.888$ である）。それなので、この 3 つの交互作用項を除いたモデルへと更新する。

```
> lm.result2 <- update(lm.result1, ~.
+ - Agriculture:Education - Examination:Catholic - Education:Catholic)
> summary(lm.result2)

Coefficients:
```

```
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                252.084973  59.495468   4.237 0.000179 ***
Agriculture                 -1.889901   0.641679  -2.945 0.005971 **
Examination                 -9.165186   3.191009  -2.872 0.007178 **
Education                    0.898174   2.982577   0.301 0.765256
Catholic                    -0.369139   0.338110  -1.092 0.283086
Infant.Mortality            -5.342534   2.612026  -2.045 0.049115 *
I(Education^2)              -0.039263   0.020120  -1.951 0.059808 .
I(Examination^2)             0.027501   0.036875   0.746 0.461229
Agriculture:Examination      0.034448   0.014430   2.387 0.023054 *
Agriculture:Catholic         0.003538   0.002372   1.491 0.145714
Agriculture:Infant.Mortality 0.049577   0.027941   1.774 0.085527 .
Examination:Education        0.092610   0.040554   2.284 0.029174 *
Examination:Infant.Mortality 0.262715   0.119639   2.196 0.035465 *
Education:Infant.Mortality  -0.117195   0.147425  -0.795 0.432502
Catholic:Infant.Mortality    0.012764   0.012953   0.985 0.331834
```

ここで Education:Infant.Mortality と Catholic:Infant.Mortality の 2 つの交互作用項を取り除いたモデルを見てみよう。

```
> lm.result3 <- update(lm.result2, ~ .
+ - Education:Infant.Mortality - Catholic:Infant.Mortality)
> summary(lm.result3)

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                236.690821  53.193661   4.450 8.78e-05 ***
Agriculture                 -2.047557   0.597358  -3.428  0.00161 **
Examination                 -6.731892   2.072904  -3.248  0.00262 **
Education                   -1.474926   0.524988  -2.809  0.00817 **
Catholic                    -0.070114   0.141614  -0.495  0.62371
Infant.Mortality            -4.833092   2.284621  -2.115  0.04179 *
I(Education^2)              -0.032127   0.015133  -2.123  0.04112 *
I(Examination^2)             0.022293   0.036147   0.617  0.54152
```

```
Agriculture:Examination        0.032234   0.013913   2.317  0.02667 *
Agriculture:Catholic           0.003053   0.002285   1.336  0.19043
Agriculture:Infant.Mortality   0.060197   0.024142   2.493  0.01768 *
Examination:Education          0.083535   0.035719   2.339  0.02537 *
Examination:Infant.Mortality   0.160592   0.064505   2.490  0.01784 *
```

どうやら、さっきから Examination^2 の項があることによって、Catholic の項が有意にならないような雰囲気である。もう少し具体的にいうと、Examination^2 をモデルに組み込むことによって、Chatholic の説明率を下げているようである、ということである。そこで Examnination^2 の項を取り除いて、再度、モデルを解析してみよう。

```
> lm.result4 <- update(lm.result3, ~ . - I(Examination^2))
> summary(lm.result4)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 228.385099  51.003026   4.478 7.68e-05 ***
Agriculture                  -1.918621   0.554603  -3.459 0.001442 **
Examination                  -5.902973   1.564075  -3.774 0.000597 ***
Education                    -1.642260   0.445441  -3.687 0.000764 ***
Catholic                     -0.031804   0.126131  -0.252 0.802400
Infant.Mortality             -4.814805   2.264116  -2.127 0.040582 *
I(Education^2)               -0.037335   0.012446  -3.000 0.004953 **
Agriculture:Examination       0.027386   0.011378   2.407 0.021499 *
Agriculture:Catholic          0.002491   0.002077   1.199 0.238453
Agriculture:Infant.Mortality  0.058042   0.023675   2.452 0.019355 *
Examination:Education         0.098905   0.025360   3.900 0.000416 ***
Examination:Infant.Mortality  0.160501   0.063931   2.511 0.016827 *
```

どうやら、Examination^2 の項が Catholic の説明率を下げているようではないようである。すると、Chatolic の説明率を下げているのは Agriculture:Catholic なのだろうか、と疑えるので Examination^2 はそのままにして、Agriculture:Catholic の交互作用項のみを取り除いてみた場合を見てみよう。

```
> lm.result4 <- update(lm.result3, ~ . - Agriculture:Catholic)
> summary(lm.result4)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 226.449866  53.225390   4.255 0.000149 ***
Agriculture                  -1.786269   0.570728  -3.130 0.003519 **
Examination                  -5.808849   1.976168  -2.939 0.005789 **
Education                    -1.425913   0.529545  -2.693 0.010802 *
Catholic                      0.113057   0.035844   3.154 0.003299 **
Infant.Mortality             -5.381088   2.272558  -2.368 0.023547 *
I(Education^2)               -0.040106   0.014060  -2.853 0.007233 **
I(Examination^2)              0.003042   0.033521   0.091 0.928215
Agriculture:Examination       0.018650   0.009603   1.942 0.060213 .
Agriculture:Infant.Mortality  0.063326   0.024296   2.606 0.013351 *
Examination:Education         0.090751   0.035702   2.542 0.015607 *
Examination:Infant.Mortality  0.176270   0.064135   2.748 0.009405 **
```

やはり、Catholic の説明率を下げていたのは Agriculture:Catholic だったようである。しかしそれでも、結局 `Examination^2` の項は不要なので取り除いてみよう。

```
> lm.result5 <- update(lm.result4, ~ . - Examination^2)
> summary(lm.result5)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 225.408202  51.251845   4.398  9.3e-05 ***
Agriculture                  -1.773153   0.544463  -3.257 0.002460 **
Examination                  -5.701934   1.564507  -3.645 0.000838 ***
Education                    -1.454521   0.419562  -3.467 0.001381 **
Catholic                      0.113804   0.034403   3.308 0.002139 **
Infant.Mortality             -5.361763   2.231173  -2.403 0.021534 *
I(Education^2)               -0.040713   0.012197  -3.338 0.001971 **
Agriculture:Examination       0.018269   0.008518   2.145 0.038783 *
```

```
Agriculture:Infant.Mortality   0.062883   0.023470   2.679 0.011054 *
Examination:Education          0.093029   0.025033   3.716 0.000684 ***
Examination:Infant.Mortality   0.175787   0.063028   2.789 0.008395 **
```

すると、全ての項が有意になったのでこれで最終的なモデルが出来上がったことになる。今回の場合、もっと初期の段階で Examination^2 の 2 次項を取り除いてもよかったかもしれない。しかし、それをしなかったのは最初の gam() によって得られた図より、曲線関係が強く疑われたためである。

3.7.2　step() を用いた解析例

以下には step() によるモデル選択の結果を示す。手順としては step.result に全ての 2 次の交互作用項を含むモデル解析の結果を代入し、それを step() の引数に指定することになる。

前節では各項の p 値をもとに変数を削除したり追加したりしていたが（**変数増減法**）、step() では AIC をモデル選択の判断基準としている。AIC は値が小さければ小さいほど、そのモデルが良いことを示すものである。この step() を用いた方法については文献 [15] に詳しい解説が載っているので、そちらも参照するとよいだろう。

```
> # 全ての 2 次の交互作用項を含むモデル
> step.result <- lm(Fertility ~ . ^2, data=swiss)
> step.result2 <- step(step.result) # step() によるモデル選択
> summary(step.result2) # 係数表の表示

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 225.910055  52.457566   4.307 0.000128 ***
Agriculture                  -1.906654   0.562824  -3.388 0.001756 **
Examination                  -5.120204   1.578305  -3.244 0.002593 **
Education                    -2.473498   1.202766  -2.057 0.047249 *
Catholic                      0.211157   0.054180   3.897 0.000420 ***
Infant.Mortality             -5.269347   2.287270  -2.304 0.027292 *
Agriculture:Examination       0.014880   0.009277   1.604 0.117706
Agriculture:Education         0.019081   0.013718   1.391 0.173013
Agriculture:Infant.Mortality  0.063530   0.024021   2.645 0.012158 *
```

```
Examination:Education          0.063891   0.030098   2.123 0.040925 *
Examination:Infant.Mortality   0.172194   0.064887   2.654 0.011891 *
Education:Catholic            -0.012381   0.005237  -2.364 0.023741 *
```

3.7.3 連続変数間の交互作用

2 要因の分散分析モデルではカテゴリカル型変数間の交互作用について考える必要があった。これは 2 つの要因のうち、一方の要因の水準における値（観測値）がもう一方の要因の水準に依存して変動するということであった。

カテゴリカル型変数間の交互作用についての考察は少し複雑で難しいものである。しかし連続変数間の交互作用の解釈はより簡単である。なぜならば、連続変数間の交互作用は 2 つの変数が乗法的に応答変数へ作用するということだからである。つまり、重回帰モデルにおける $X_1 \times X_2$ という交互作用項は、数学そのままの「かける」（乗法）という意味なのである。

3.7.4 回帰診断

回帰分析を行ったならば、その分析が適切なものであったかどうかを診断する必要があり、この作業を**回帰診断**という。この回帰診断では (1) 分散の均一性、(2) 誤差の正規性の 2 つについて考えることになる。これは `lm()` による解析結果が代入されているオブジェクトを `plot()` の引数に指定することで、回帰診断のために必要な図が出力される。

試しに、先ほどの解析結果が格納されているオブジェクト `step.result2` より、回帰診断図を出力してみよう（図 3.11）。`plot()` の引数に `which=1:4` が指定されているが、これは出力される診断図の種類を指定しているのである。

```
> windows()
> par(mfrow=c(2, 2))
> plot(step.result2, which=1:4)
```

分散の均一性については、標準化された残差をプロットする（図 3.11 の左下の図）ことによって確認することができる。図の見方としては、データポイントが全体に満遍なく散布していることが望ましい。適合値（Fitted values）が大きくなるにつれて分散も大きくなるような傾向を示す（扇のような形になる）ようだとまずいわけである。

続いて誤差の正規性については Q-Q プロットとよばれるものを用いればよい（図 3.11 の

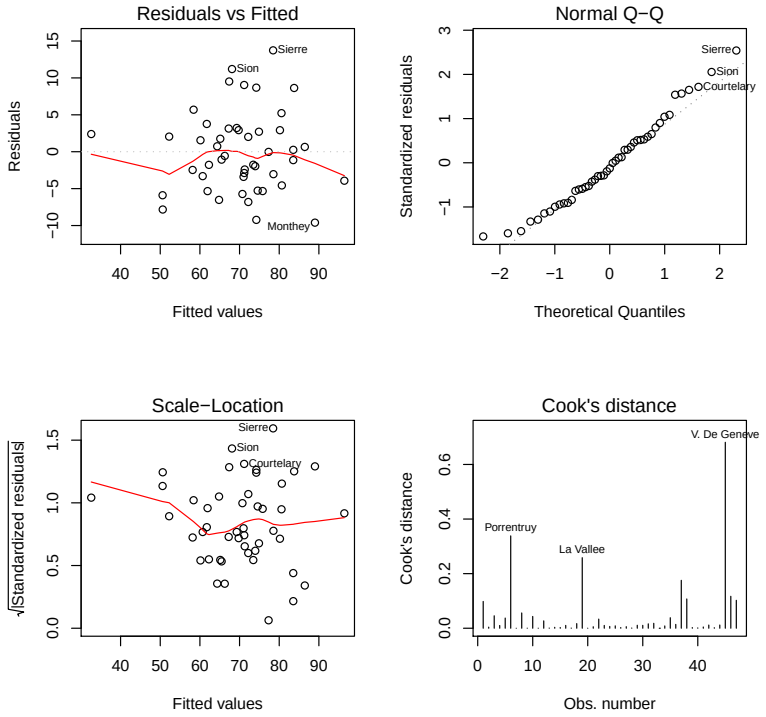


図 3.11　step.result2 の回帰診断図

右上の図)。これは描かれたデータポイントが直線にぴったりと乗っている状態が最も好ましい。あまりにもデータポイントが直線から離れている（全体的に曲線的に見える）ようだと正規性に疑いがあると判断できる。

ちなみに、左上の図は残差プロットである。これは実測値と予測値の差（残差）を示している。そして右下の図ははずれ値を見つけるためのものだと思えばよいだろう。この図において、特別に突起した線があるようだと、その番号のデータははずれ値であり、結果に良くない影響を及ぼしているかもしれないと判断するわけである。

3.8　共分散分析モデル

ここまで説明変数がカテゴリカル型のみである分散分析モデルと、説明変数が連続型のみである回帰モデルを扱ってきた。一般線形モデルにはこの 2 つのモデルに加え、カテゴリカル型と連続型を混在させた**共分散分析モデル**も扱うことができる。

古典的な教科書では共分散分析は 2 本の平行な回帰直線が当てはめられる分析法であり、事前に 2 本の回帰直線の平行性の検定を行うべし、などと解説されていることがある。これはカテゴリカル型と連続型の交互作用項が有意であるかどうかを確かめることにあたる。つまり、もし交互作用が認められるならば、当てはめられる 2 本の回帰直線は平行ではなくなるし、交互作用が認められないならば、2 本の直線は平行になるということである。

そこで、この章ではカテゴリカル型と連続型の 2 変数に交互作用が認められない場合の

例と交互作用が認められる場合の例を紹介する。前者は当てはめられる回帰直線が平行になり、後者は平行ではなくなることを確認してもらいたい。

3.8.1 加法モデル

交互作用項が有意でない場合、それは加法モデルである。共分散分析において、これは当てはめられる回帰直線の傾きが同じであることを意味している。例えば、表 3.9 のデータセットにおいて、数学（Math.）と物理（Physics）相関関係を性別（Sex）を考慮して考えてみよう。モデル式で表現すると式 3.11 のように書くことができる。

$$Physics = Math. + Sex \tag{3.11}$$

表 3.9 4 教科の期末テスト成績

Math.	Physics	Japanese	English	Sex
87	90	69	50	Male
70	72	59	60	Male
85	83	75	57	Male
60	55	43	51	Male
96	100	86	75	Male
41	58	53	56	Male
60	55	69	70	Male
50	60	65	59	Male
79	88	80	70	Male
87	72	85	90	Female
68	54	67	74	Female
84	65	82	88	Female
59	37	59	68	Female
96	82	89	99	Female
40	40	38	40	Female
59	37	57	76	Female
48	42	46	57	Female
78	70	77	87	Female

このモデル式（式 3.11）は、「$Math.$ と $Physics$ の散布図へ Sex の水準（$Male, Female$）ごとに回帰直線を当てはめたとき、それぞれの回帰方程式の切片（定数項）と傾きは有意に異なるか」ということを考えている。これについて、実際の解析結果をもとに詳しく解説していこう。

```
> mydata
   Math Physics Japanese English    Sex
1    87      90       69      50   Male
2    70      72       59      60   Male
3    85      83       75      57   Male
（以下省略）

> attach(mydata)
> ancova.add.1 <- lm(Physics ~ Math * Sex)
> summary(ancova.add.1)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     13.43681   10.12727   1.327    0.206
Math             0.85998    0.14084   6.106 2.71e-05 ***
SexFemale      -15.83789   14.07415  -1.125    0.279
Math:SexFemale  -0.01893    0.19689  -0.096    0.925

> ancova.model.2 <- update(ancova.add.1, ~. -Math:Sex)
> summary(ancova.add.2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.11281    7.04477   2.003 0.063555 .
Math          0.85029    0.09511   8.940 2.14e-07 ***
SexFemale   -17.14971    3.34255  -5.131 0.000123 ***

> dummy.coef(ancova.add.2)
Full coefficients are

(Intercept):     14.11281
Math:           0.8502941
Sex:                 Male    Female
                  0.00000 -17.14971
```

最初に交互作用項を含めたフルモデルを解析してみた（ancova.add.1）。しかし交互作用項（Math:SexFemale）は有意でないことがわかる（$p = 0.925$）。そこで update() を用いて、交互作用項を除いた加法モデルへとモデルを更新した（ancova.add.2）。すると今度は Math と SexFemale が有意となったので（それぞれ $p = 0.0000, p = 0.0001$）、どうやら加法モデルを当てはめることが最適であるようだということがわかる。

加法モデルである場合、図 3.12 に示すように、当てはめられる 2 本の回帰直線は平行となる。つまり、傾きは同じだが切片のみが異なるということである。これは dummy.coef(ancova.add.2) の出力を読み取ることができると納得できるはずである。

より分かりやすいように数式でまとめてみると式 3.12 のように書ける。これはまさにカテゴリカル型変数の水準（この場合は男性と女性）によって直線の切片（定数項）が変化することを意味している。男性の場合だと切片は 14.11 であり、女性の場合だと $14.11 - 17.15 = -3.04$ ということになる。

$$Physics = 14.11 + 0.85 \times Math. + \left\{ \begin{array}{c} 0.00 \times Male \\ -17.15 \times Female \end{array} \right\} \tag{3.12}$$

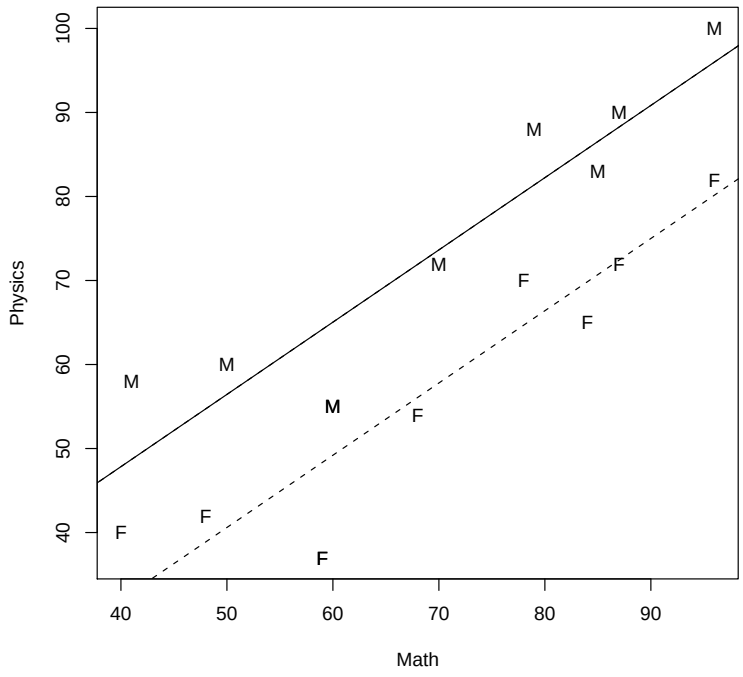


図 3.12　加法モデルが当てはめられた場合

選ばれたモデルは本当に正しいか

共分散分析モデルを扱う際には、分散分析モデルや重回帰モデルよりも注意深く交互作用について考察しなければならない。もし勘のよい読者ならば今回の解析結果に対して疑問を

もったことだろう（もし疑問をもてたならば、あなたは相当にデータ解析の技術が備わってきていると考えてよい)。

実は summary(ancova.add.2) の出力を見ると、切片（Intercept）は 5% 水準で有意ではないことがわかる（$p = 0.0636$)。これは切片が 0 であるということを意味しているが、理論的にテストの得点は 0 以下にはならないので切片が 0 であるということは妥当な結果であるといえる。そして切片が 0 であるということは、当てはめられる 2 本の回帰直線は完全に一致するということも意味している。傾きが水準によって依存しないのであれば、こうなるのは容易に想像できるだろう。

まとめると、共分散分析モデルにおいて、理論的に切片が 0 であると考えた方が妥当な場合、交互作用が有意でなければ当てはめられる回帰直線は完全に一致する。このような場合、交互作用が有意でないと確認できた時点で群間差は認められないと考えるのが妥当である。

しかし、理論的には切片が 0 であると考えられても、実質的にそうではないモデルの方が適切であると考えるならば、無理に切片を 0 であるとする必要はない。例えば今回のデータセットより、$Physics$ と $Math$ の 95% の信頼区間を求めてみると、それぞれ $[54.93 73.96], [60.29 78.27]$ となる。つまり、実質的に 0 点を取る学生はいないと考えてもよいのである。

こうした考察は単にモデル解析の結果（特に p 値）にのみ着目してしまうと難しい。モデル解析には事前にデータをプロットして全体の散布度を確認したり、今回のように平均の信頼区間を求めてみたりなど、総合的な視点からデータと向き合わなければならないのである。

3.8.2 交互作用モデル

今度は表 3.9 の $Japanese$ と $English$ を用いて交互作用項が有意である場合の解析例を示す。今回解析しようとするモデルは式 3.13 のように書ける。

$$English = Japanese + Sex \tag{3.13}$$

```
> ancova.intr.1 <- lm(English ~ Japanese * Sex)
> summary(ancova.intr.1)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        29.9936    10.6455   2.817   0.0137 *
Japanese            0.4642     0.1571   2.954   0.0105 *
SexFemale         -20.5779    13.3636  -1.540   0.1459
Japanese:SexFemale  0.5262     0.1962   2.682   0.0179 *
```

```
> dummy.coef(ancova.intr.1)
Full coefficients are

(Intercept):      29.99360
Japanese:        0.4642031
Sex:                  Male    Female
                   0.00000 -20.57794
Japanese:Sex:         Male    Female
                  0.0000000 0.5262288
```

さて summary(ancova.intr.1) の出力結果を見ると交互作用項は有意である（$p = 0.0179$）。ここで SexFemale の項が有意ではない（$p = 0.1459$）が、これは気にする必要はない。なぜなら 2 要因の分散分析モデルの交互作用の解説でも述べたように、交互作用は主効果による説明をも含んでいるからである。つまり交互作用が有意に認められるならば、主効果として働いている変数も必ず意味のある変数であるとして考えるべきなのである。

したがって、交互作用を含むフルモデルがこの場合では適切なモデルであるといえる。これを数式でまとめなおすと式 3.14 のようになる。これは切片が水準に依存することと、さらに交互作用があることで傾き（$Jap.$）も水準に依存して変化することを意味している（図 3.13）。

$$English = 29.99 + 0.46 \times Jap. + \left\{ \begin{array}{c} 0.00 \times Male \\ -20.58 \times Female \end{array} \right\} + \left\{ \begin{array}{c} 0.00 \times Jap.(Male) \\ 0.53 \times Jap.(Female) \end{array} \right\} \quad (3.14)$$

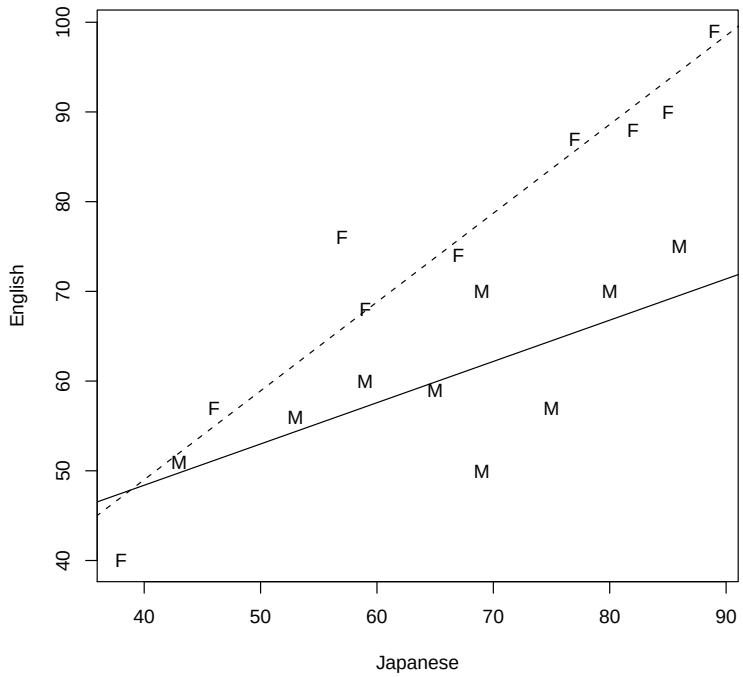


図 3.13　交互作用モデルが当てはめられた場合

第 4 章

一般化線形モデル

一般線形モデルには分散分析モデル、回帰分析モデル、共分散分析モデルが含まれるということを示した。また、これらのモデル解析において、応答変数は連続型であった。これ以降は一般化線形モデルといって、今までやってきた一般線形モデルを拡張したモデル解析の理論を紹介する。この一般化線形モデルには計数データの解析に用いられるポアソン回帰モデル、2 値データの解析に用いられるロジスティック回帰モデルが含まれる。

4.1 ポアソン回帰モデル

4.1.1 古典的なアプローチ

まず最初に**分割表の解析**について、古典的な χ^2 **検定（独立性の検定）**を用いた解析について見てみよう。

表 4.1 は未婚の大学生に対して「将来、結婚したいと思うか」を Yes か No の 2 件法で、「将来、子供が欲しいと思うか」を Yes か No の 2 件法で聞いた結果をまとめたものである。表のデータを見る限り、どうやら結婚を望んでいる人は、同時に子供も欲しいと考えているようであることがわかる (marry-Yes:children-Yes = 124)。

表 4.1　結婚と子供に対する望み

	children	
marry	Yes	No
Yes	124	6
No	12	9

独立性の検定とは表 4.1 のような分割表に対して、行要因（$marry$）と列要因（$children$）に関連があるかどうかを確かめるための方法である。この検定における帰無仮説と対立仮説は次のように立てられる。

- H_0 : 行要因と列要因は互いに独立である（関連がない）
- H_1 : 行要因と列要因は互いに独立ではない（関連がある）

R には chisq.test() という関数が用意されているので、これを利用して独立性の検定を行ってみよう。

```
> dat <- matrix(c(124, 6, 12, 8), ncol=2, byrow=TRUE)
> dat
     [,1] [,2]
[1,]  124    6
[2,]   12    8
> chisq.test(dat)

        Pearson's Chi-squared test with Yates' continuity correction

data:  dat
X-squared = 21.6354, df = 1, p-value = 3.297e-06

Warning message:
In chisq.test(dat) :  カイ自乗近似は不正確かもしれません
```

エラーメッセージが出るが、これは気にしなくてよい。これによれば $p = 0.0000(3.297e-06 = 0.000003297)$ であるから 5% 水準で有意である。したがって帰無仮説は棄却され、行要因と列要因は独立ではないといえる。つまり $marry$ と $children$ には関連があるということである。

ただし、この解析では相関係数のように 2 変数間にどの程度の強さの関連があるのかは分からない。そのようなときに役立つのが**連関係数**と呼ばれる指標である。連関係数には四分点位相関係数やユールの連関係数というものがあるが、詳しくは文献 [18] に譲ることにする。

4.1.2　ポアソン回帰による分割表の解析

前述してきたような古典的な方法（独立性の検定）を用いることと同義である、別のアプローチが**ポアソン回帰モデル**による解析である。

ポアソン回帰モデルとは応答変数が**計数データ**である場合に用いられるべきモデルである。計数データとは「1 つ、2 つ、3 つ、・・・」といったように飛び飛びの値をとるような

データで、数学では分離量、あるいは離散量と呼ばれている。説明変数にはカテゴリカル型の変数がおかれることになる。もちろん、説明変数に連続型をおくこともできるし、カテゴリカル型と連続型を混在させることも可能である。

それでは、早速だが実際に表 4.1 のデータセットをポアソン回帰モデルとして解析してみよう。一般線形モデルの解析には lm() を用いたが、一般化線形モデルでは glm() を使うことになる。使い方はほぼ同じだが、ポアソン回帰モデルである場合には、引数に family=poisson と指定する点が異なる。

```
> count <- c(124, 6, 12, 8) # セル内の計数データ

> # 2 水準のカテゴリカル型データを用意する
> marry <- c(1, 1, 2, 2)
> marry <- factor(marry, levels=c(1, 2), labels=c("Yes", "No"))

> # 中身の確認: 第 1 水準は Yes, 第 2 水準は No である
> marry
[1] Yes Yes No  No
Levels: Yes No

> #以下同様
> children <- c(1, 2, 1, 2)
> children <- factor(children, levels=c(1, 2), labels=c("Yes", "No"))
> children
[1] Yes No  Yes No
Levels: Yes No

> # モデルの指定法は lm() と全く同じ
> # 引数に family=poisson と指定する
> result <- glm(count ~ marry * children, family=poisson)

> # 係数表の出力
> summary(result)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)          4.8203     0.0898  53.676  < 2e-16 ***
marryNo             -2.3354     0.3023  -7.725 1.12e-14 ***
childrenNo          -3.0285     0.4180  -7.245 4.32e-13 ***
marryNo:childrenNo   2.6231     0.6189   4.238 2.25e-05 ***

Residual deviance: 1.7764e-15  on 0  degrees of freedom
```

出力される係数表の形式はほぼ同じだが、復習の意味もかねて 1 つずつ説明していこう。Estimate は係数（ポアソン回帰係数とでもいうべきか）である。Std. Error は標準誤差であり、これは係数の推定精度を表しているのであった。標準誤差は小さいほど、より正確な推定ができているといえる。続いて z value は検定統計量の z 値である。これは lm() で得られる t value にあたるものである。最後に Pr(>|z|) は p 値である。lm() では検定統計量に t 値が用いられていたが、glm() では（ポアソン回帰では）z 値が用いられるという点で異なっているのである。

また、新たに Residual deviance という出力が得られていることに気がついただろう。これは**過分散**の疑いがあるかどうかを判断するためのものである。残差逸脱度（1.7764e-15）が自由度（0）より大きいときに過分散の疑いがある。この過分散とは、簡単にいえば「モデルに含まれる説明変数だけでは応答変数の変動を説明できていない」という事実を示している。もっと単純にいってしまえば、予測の役に立たない説明変数ばかりがモデルに含まれているということである。

今回の解析結果では残差逸脱度が自由度より大きいとはいえないので、過分散の疑いは薄いといえるだろう。つまり、モデルに含まれている説明変数（$marry$ と $children$）によって、十分によい予測ができていることになる（何を予測するのかといえば、もちろんセル内の計数データである）。

さて、数値の見方を説明し終えたところで、今度は結果の解釈について解説しよう。分割表の解析において重要なのは「交互作用項が有意であるかどうか」ということである。したがって、主効果（単一の変数による影響）について考えることに意味はない。なぜなら、ポアソン回帰において交互作用が認められるかどうかについて検定することは、独立性の検定をすることと同義だからである。

係数表を見ると marryNo:childrenNo の項は有意であることが確認できる（$p = 0.0000225$）。2 要因の分散分析において交互作用が認められるということは、応答変数の変動が 2 変数のある水準の組合せに大きく依存するということであった。これはポアソン回帰の場合も同じである。

つまり、結婚したい（$marryYes$）という水準と子供が欲しい（$childrenYes$）という水準の組合せによって、応答変数である $count$ のデータが大きく変動するということである。表

4.1 を見ても、この 2 つの水準の組合せの元での度数が顕著に大きいということが分かるだろう（marry-Yes:children-Yes = 124）。もしも、各セルの度数がほぼ等しい状態であったならば、交互作用項は有意ではなくなるはずである。

例えば、各セルに入力されているデータが全て等しい場合を解析してみれば、交互作用が有意でないということはどういうことなのかよく分かるだろう。

```
> count2 <- c(50, 50, 50, 50)
> summary(glm(count2 ~ marry * children, family=poisson))

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.9120     0.1414   27.66   <2e-16 ***
marryNo             0.0000     0.2000    0.00        1
childrenNo          0.0000     0.2000    0.00        1
marryNo:childrenNo  0.0000     0.2828    0.00        1

Residual deviance: -1.2434e-14  on 0  degrees of freedom
```

ところで、今回の解析で得られた係数をモデル式に当てはめれば、各セルの観測値が予測できるはずだと考えるだろう。試しに `dummy.coef()` を利用してダミー係数式を出力し、それを数式でまとめてみると式 4.1 のようになる。

$$count = 4.82 + \left\{ \begin{matrix} 0.00 \times Yes(marry) \\ -2.34 \times No(marry) \end{matrix} \right\} + \left\{ \begin{matrix} 0.00 \times Yes(children) \\ -3.02 \times No(children) \end{matrix} \right\} + \left\{ \begin{matrix} 0.00 \times Yes : Yes(marry : children) \\ 0.00 \times No : Yes(marry : children) \\ 0.00 \times Yes : No(marry : children) \\ 2.62 \times No : No(marry : children) \end{matrix} \right\} \tag{4.1}$$

例えば 1 行 1 列目にあるセルの度数を予測したいとしよう。つまり $marryYes$ と $childrenYes$ のセルの度数を予測するということである。計算してみると $4.82 + 0.00 + 0.00 + 0.00 = 4.82$ となるが、これは表 4.1 にある 124 という度数とは明らかに違う。いったいどういうことだろうか。

実はこの例で解析されたモデルは**対数線形モデル**とも呼ばれ、対数線形モデルは式 4.2 のように表される。λ は定数項、λ_i^X は変数 X（i は水準数）、λ_j^Y は変数 Y（j は水準数）、λ_{ij}^{XY} は X と Y の交互作用項である。

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \tag{4.2}$$

さて、式 4.2 左辺の左辺（応答変数）を見ると、観測値の予測値（μ_{ij}）は対数で表されていることが分かるだろう。つまり、式 4.1 より得られた 4.82 という数値は予測値の対数を取ったものなのである。したがって、各セルの予測値を得るためには e（これはネイピア数と呼ばれるものである）を底とする数式のべき乗を計算しなければならない。

難しそうなことをいったが、要は R にある `exp()` という関数の引数に 4.82 という値を指定すれば計算できる。もっとも、そのような面倒なことをせずとも `predict()` や `fitted()` といった便利な関数が用意されているので、それを使えばよい。ただし、最初の数回は練習のために自分で計算してみるとよいだろう。

```
> # ダミー係数式より予測値を得る
> predict(result)
       1        2        3        4
4.820282 1.791759 2.484907 2.079442

> # これを exp() の引数に指定すれば、各セルの予測値が得られる
> exp(predict(result))
  1   2   3   4
124   6  12   8

> # fitted() を使えば一気にできてしまう
> fitted(result)
  1   2   3   4
124   6  12   8
```

4.2　ロジスティック回帰モデル

ロジスティック回帰分析といっても、その用途は非常に広いものである。またロジスティック回帰モデルは前章で解説したポアソン回帰モデルとも似ている手法である。実際、表 4.1 のデータセットをロジスティック回帰モデルとして解析することも可能である。

ところでロジスティック回帰分析とは、これ以降に紹介する**比率データ**へのロジスティック曲線の当てはめ、2 値ロジットモデル、名義・順序ロジットモデルといったモデルの総称である。また、ロジスティック回帰分析はしばしばロジット分析と呼ばれたりする。

そこでまずは最も典型的なロジスティック回帰分析の一例でもある、比率データへのロジ

スティック曲線の当てはめを紹介する。この方法が薬の致死量（LD-50）を算出するために利用されている、というのはよく知られた例である。

4.2.1　致死量の算出

マウスに投与量を変えて毒性のある薬物を投与する場面を考えてみよう。例えば、10 匹のマウスに 0.1mg/kg（これは体重 1kg あたり 0.1mg の薬を投与するという意味である）の薬物を投与したとき、3 匹のマウスが死亡したとしよう。この場合の死亡率は $3/10 = 0.3$ である。投与量を増やしていけば死亡率は増えていくだろうと予想できるが、このような死亡率は**ロジスティック曲線**を当てはめることによって上手く説明できることが多い。

ロジスティック曲線とは式 4.3 のような関数が描く曲線のことである。比率データ（この場合は死亡率）にロジスティック曲線を当てはめるというのは、単回帰分析モデルで $f(x) = \alpha + \beta x$ という直線を当てはめた作業と同じである。

$$f(x) = \frac{1}{1 + \exp(\alpha - \beta)} \tag{4.3}$$

それでは表 4.2 のデータセットを用いて、実際にロジスティック曲線を当てはめ、致死量（LD-50）を算出してみよう。ポアソン回帰のときと同じく、`glm()` を用いればよいが、引数に `family=binomial` と指定することを忘れてはならない。また、応答変数には 1 列目に死亡数（$Death$）、2 列目に生存数（$Surv$）として行列を指定しなければならない。死亡率を求めるので 1 列目に死亡数をおくが、もしも生存率を求めるのであれば 1 列目に生存率をおくようにする。

表 4.2　投与量と死亡数のデータセット

投与量	死亡数	生存数	マウスの数	死亡率
5.42	3	53	56	0.05
5.61	10	47	57	0.18
5.78	15	44	59	0.25
5.95	25	28	53	0.47
6.12	49	11	60	0.82
6.28	50	6	56	0.89
6.43	58	1	59	0.98
6.58	57	0	57	1.00

```
> # 投与量
```

```
> Dose <- c(5.42, 5.61, 5.78, 5.95, 6.12, 6.28, 6.43, 6.58)
> # 死亡率
> DeathRate <- c(0.05, 0.18, 0.25, 0.47, 0.82, 0.89, 0.98, 1.00)

> # 横軸に投与量、縦軸に死亡率をとってプロット
> plot(Dose, DeathRate, cex=1.5, ylim=c(0, 1))

> Death <- c(3, 10, 15, 25, 49, 50, 58, 57) # 死亡数
> Surv <- c(53, 47, 44, 28, 11, 6, 1, 0) # 生存数
> Y <- cbind(Death, Surv) # 1 列目に死亡数、2 列目に生存数
> Y
     Death Surv
[1,]     3   53
[2,]    10   47
[3,]    15   44
[4,]    25   28
[5,]    49   11
[6,]    50    6
[7,]    58    1
[8,]    57    0

> result <- glm(Y ~ Dose, family=binomial)
> summary(result)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -37.8887     3.3053  -11.46   <2e-16 ***
Dose          6.4031     0.5561   11.52   <2e-16 ***

Residual deviance:   6.3721  on 6  degrees of freedom
```

ここで今更かと思うかもしれないが、今回のモデル式を見てみよう（式 4.4)。これは応答変数が**比率データ**であることを除けば、単回帰モデルと全く同じである。つまり、1 つの連続型の説明変数がおかれているということである。

$$DeathRate = Dose \tag{4.4}$$

したがって、予測値を求めるための式も非常に単純なものとなる（式 4.5)。例えば、$Dose = 5.42$ のときには $-37.89 + 6.40 * 5.42 = -3.202$ となる。

$$DeathRate = -37.89 + 6.40 \times Dose \tag{4.5}$$

これは `predict()` を用いれば簡単に得られる。

```
> predict(result)
         1          2          3          4          5
-3.1838350 -1.9672428 -0.8787129  0.2098169  1.2983468

         6          7          8
 2.3228455  3.2833130  4.2437805
```

しかし、これは死亡率の予測値ではない、実は式 4.3 の式を数学的に変形することによって、$y = \alpha + \beta x$ という線形モデルとして表すことができるのである。本来、ロジスティック曲線は非線形モデルなのだが、このような変換を施すことによって線形モデルとして扱うことができるのである（だから、内部的にはそのようなことが行われているということである)。詳しくは文献 [2] や文献 [4] を参照するとよいだろう。

そのようなわけで、死亡率の予測値を得るためには、式 4.5 に推定された切片（$\alpha = -37.89$）と係数（$\beta = 6.40$）を代入するのではなく、式 4.3 に代入しなければならないのである。

ポアソン回帰の章でも述べたが、最初の何回かは練習のために自分で計算するのも良いだろう。しかし、手っ取り早く死亡率の予測値を得るためには `fitted()` を利用するといい。

```
> fitted(result)
         1          2          3          4          5
0.03977859 0.12268535 0.29344456 0.55226264 0.78555662

         6          7          8
0.91075150 0.96385189 0.98584987
```

結局、ロジスティック曲線を死亡率のデータに当てはめると図 4.1 のようになる。ロジスティック曲線について、S 字カーブのちょうど真中の座標は $(x, y) = (1/2, \alpha/\beta)$ となることを覚えておくと便利である。これより $(1/2, 37.889/6.403) = (0.5, 5.917382)$ と計算できるから、$LD_{50} = 5.917$ といえる。つまり、約 6mg/kg で半数のマウスが死に至るということ

である。

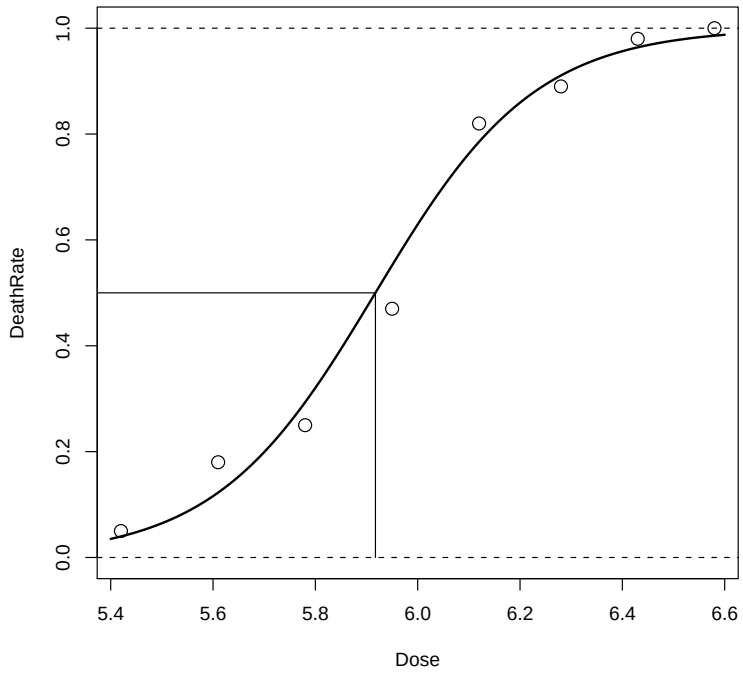


図 4.1　ロジスティック曲線の当てはめ

```
> plot(Dose, DeathRate, cex=1.5, ylim=c(0, 1))
> abline(h=c(0, 1), lty=2)
> x.jiku <- seq(5.4, 6.6, 0.01)
> y.jiku <- 1 / (1 + exp(37.889 - 6.403 * x.jiku))
> points(x.jiku, y.jiku, type="l", lwd=2)
> lines(c(0, 5.917382), c(0.5, 0.5))
> lines(c(5.917382, 5.917382), c(0.5, 0))
```

4.3　2 値ロジットモデル

さて前節では比率データへのロジスティック曲線の当てはめを紹介した。今度は 2 **値データ**を応答変数とする 2 **値ロジットモデル**を紹介する。

2 値データとは 2 つの水準を持つカテゴリカル型データのことである。例えば、「生か死か」、「男か女か」、「成功か失敗か」、「買うか買わないか」といったものがあげられる。この

ような 2 値データを応答変数とする場合、次のようなモデルが考えられるだろう。

説明変数が 1 つで連続型　一般線形モデルでは単回帰モデルとして扱われた
説明変数が 2 つ以上で連続型　一般線形モデルでは重回帰モデルとして扱われた
説明変数が 1 つでカテゴリカル型　一般線形モデルでは 1 要因の分散分析モデルとして扱われた
説明変数が 2 つ以上でカテゴリカル型　一般線形モデルでは 2 要因の分散分析モデルとして扱われた
説明変数とカテゴリカル型が混在　一般線形モデルでは共分散分析モデルとして扱われた

ただし、ここで注意しておくべき重要なことがある。それは「説明変数がカテゴリカル型のみのモデルを解析しても得ることは何もない」ということである。だから、一般線形モデルでは分散分析モデルとして扱われたようなモデルは、応答変数が 2 値データのときには用いるべきではない。そのような場合には 2 値ロジットモデルではなく、分割表に集計してポアソン回帰モデルとして解析するべきでなのである。

4.3.1　2 値ロジットモデルの解析例

表 4.3 はあるコンビニエンスストアにおいて、来客者がアイスを買うか否かについて取られたデータであり、このデータセットには $Purchase$（購買行動: Yes, No）、Sex（性別: Female, Male）、$Weather$（天候: Cloudy, Rainy, Sunny）、$Discount$（値引き額）、$Temp$（気温）の変数が含まれている。ここで扱いたい問題は Sex、$Time$、$Discount$、$Temp$ といった変数で $Purchase$ を予測できるかということである。違う言い方をすれば「アイスの購買行動に影響する要因は何か」ということである。

このデータ分析には少しばかり気を使う必要がある。本来は最も複雑なモデル（フルモデル）から解析していくが、全ての交互作用項を含めたモデルでは推定するパラメータが多すぎてうまく解析できない。そこでまずは 4 つ全ての変数を用いた加法モデルとして解析を行う。

なお以下に示す解析例において、表 4.3 のデータセットが `dat` に格納されているものとする。また、`read.table()` などを利用してデータを読み込む場合、カテゴリカル型変数の水準の順序について確認しておいた方がよい。これは `levels()` の引数に対象となるオブジェクト（変数）を指定することによって確認できる。

```
> dat <- read.table("logit.dat", header=TRUE)
> attach(dat)
> levels(Sex)
[1] "Female" "Male"
```

表 4.3　購買行動のデータセット

Purchase	Sex	Weather	Discount	Temp
Yes	Male	Sunny	20	33
Yes	Male	Sunny	20	30
No	Female	Cloudy	0	25
Yes	Male	Cloudy	0	29
No	Male	Rainy	0	22
No	Female	Cloudy	0	25
No	Female	Sunny	10	29
Yes	Male	Sunny	0	28
Yes	Female	Rainy	0	20
Yes	Male	Sunny	20	35
Yes	Male	Sunny	20	35
Yes	Male	Sunny	20	31
No	Male	Cloudy	10	26
Yes	Female	Sunny	0	33
Yes	Male	Sunny	10	31
No	Female	Cloudy	10	26
No	Female	Rainy	10	23
No	Male	Sunny	0	28
Yes	Male	Sunny	20	30
No	Female	Sunny	20	29
No	Female	Cloudy	0	29
Yes	Male	Cloudy	10	25
No	Female	Cloudy	10	28
Yes	Male	Cloudy	20	30
Yes	Male	Sunny	10	33
Yes	Male	Sunny	0	35
No	Female	Sunny	0	28
No	Male	Rainy	0	25
No	Female	Rainy	0	23
Yes	Male	Sunny	0	36
Yes	Male	Sunny	20	30
No	Female	Cloudy	20	26
Yes	Male	Sunny	10	29

```
> levels(Weather)
[1] "Cloudy" "Rainy"  "Sunny"
---
> summary(result)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -13.01151    7.62518  -1.706   0.0879 .
SexMale        2.54978    1.08062   2.360   0.0183 *
WeatherRainy   1.04576    1.96881   0.531   0.5953
WeatherSunny   0.41916    1.37961   0.304   0.7613
Discount       0.03408    0.07022   0.485   0.6275
Temp           0.38898    0.27640   1.407   0.1593
```

出力された係数表を見ると、$Weather$ と $Discount$ は明らかに有意ではないことが分かる。それなので、この 2 つの変数を除いたモデルへと更新する。

```
> result2 <- update(result, ~ . -Weather -Discount, data=dat)
> summary(result2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3155     5.1165  -2.407   0.0161 *
SexMale       2.6633     1.0665   2.497   0.0125 *
Temp          0.3837     0.1770   2.168   0.0302 *
```

すると今度は Sex と $Temp$ の項が有意になった。どうやらこの 2 つの変数がアイスの購買行動に影響しているようであるが、この 2 つの変数の交互作用はどうであるかを確認してみる。改めて Sex と $Temp$ の交互作用項を組み込んで解析してみる。

```
> result3 <- update(result2, ~ . ^2, data=dat)
> summary(result3)
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.818790   6.316914  -0.288    0.773
SexMale      -20.892266  13.094251  -1.596    0.111
Temp           0.004307   0.236665   0.018    0.985
SexMale:Temp   0.853688   0.481040   1.775    0.076 .
```

Sex と $Temp$ の交互作用は有意に認められない（$p = 0.076$）ので、今回の解析では交互作用項を含めないモデルの方が適切であるといえそうである。交互作用を含めたモデルと、交互作用を取り除いたモデルのどちらがより適切であるかを統計的に判断するためには anova() を利用するとよい。これは「単純化されたモデルの方が最適である」という帰無仮説を検定するためのもである。換言すると「変数を取り除くことによって、説明力は低くならない」といえる。

```
> anova(result3, result2, test="Chi")
Analysis of Deviance Table

Model 1: Purchase ~ Sex + Temp + Sex:Temp
Model 2: Purchase ~ Sex + Temp
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        29    20.2377
2        30    24.5841 -1  -4.3464    0.0371
```

この結果によれば「単純化されたモデルの方が最適である」という帰無仮説は棄却される。したがって、単純化される前のモデル（つまり交互作用を含んだモデル）がより適切なモデルであるといえる。

おそらく今回の例において交互作用項が有意ではなかったのは、データ数が少なかったためであると考えられる。いずれにしても、ここは練習として交互作用モデルと加法モデルの両方について考えてみよう。

交互作用モデルの解釈

交互作用モデルの係数を数式でまとめると式 4.6 のように書ける。定数項の -1.82 という値は Sex によって変動し、$Temp$ の係数である 0.004 という値は $Sex : Temp$ の交互作用に

よって変動することに注意しよう。

$$Purchase = -1.82 + 0.004 \times Temp + \left\{ \begin{matrix} 0.00 \times Female \\ -20.89 \times Male \end{matrix} \right\} + \left\{ \begin{matrix} 0.00 \times Female : Temp \\ 0.85 \times Male : Temp \end{matrix} \right\} \tag{4.6}$$

女性の予測値は式 4.7 によって求められる。また、女性がアイスを買う確率は式 4.8 より得られる。これは式 4.9 に定数項 α と係数 β を代入したものである。

$$Purchase(Female) = -1.82 + 0.004 \times Temp \tag{4.7}$$

$$Purchase(Female) = \frac{exp(-1.84 + 0.004 \times Temp)}{1 + exp(-1.84 + 0.004 \times Temp)} \tag{4.8}$$

$$y = f(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)} \tag{4.9}$$

同様にして男性の場合は次のようになる。

$$Purchase(Male) = (-1.82 - 20.89) + (0.004 + 0.85) \times Temp \tag{4.10}$$

$$Purchase(Male) = \frac{exp(-22.71 + 0.854 \times Temp)}{1 + exp(-22.71 + 0.854 \times Temp)} \tag{4.11}$$

これらは関数 `predict()`、あるいは `fitted` を利用すれば簡単に計算することができる。この手順を以下に、横軸に $Temp$、縦軸に購入確率の予測値をプロットしたものと共に示す（図 4.2）。図 4.2 を見ると、交互作用モデルでは気温が 30 度近くになると男性はアイスを買う確率が高くなるが、女性は 30 度を越してもアイスを買う確率に変化はないことがわかる。

```
> dummy.coef(result3) # ダミー係数式の表示
Full coefficients are

(Intercept):       -1.81879
Sex:                 Female      Male
                    0.00000 -20.89227
Temp:           0.004307159
Sex:Temp:            Female      Male
                  0.0000000 0.8536877
```

```
> alpha <- -1.81879 + 0.00000 # 女性のα
> beta <- 0.004307159 + 0.0000000 # 女性のβ

> Ff <- function(x) # 女性の購買確率を返す関数を定義
+ exp(alpha + beta * x) / (1 + exp(alpha + beta * x))

> # fitted() で得られる女性の購買確率
> fitted(result3)[Sex=="Female"]
        3         6         7         9        14        16
0.1530197 0.1530197 0.1552660 0.1502493 0.1575391 0.1535787
       17        20        21        23        27        29
0.1519066 0.1552660 0.1552660 0.1547019 0.1547019 0.1519066
       32
0.1535787

> dat$Temp[dat$Sex=="Female"] # 女性に該当する気温データ
 [1] 25 25 29 20 33 26 23 29 29 28 28 23 26
> # 先ほど定義した関数を用いて購買確率を求める
> Ff(dat$Temp[dat$Sex=="Female"])
 [1] 0.1530197 0.1530197 0.1552660 0.1502493 0.1575391
 [6] 0.1535787 0.1519065 0.1552660 0.1552660 0.1547019
 [11] 0.1547019 0.1519065 0.1535787

> 購買行動の予測確率をプロット
> plot(Temp, fitted(result3), pch=c("F", "M")[Sex])
> x.jiku <- seq(15, 40, 0.1) # x 軸のデータ
> Fy <- Ff(x.jiku) # y 軸のデータ
> points(x.jiku, Fy, type="l") # ロジスティック曲線を描く

> # 以下、男性も同様
> alpha <- -1.81879 - 20.89227
> beta <- 0.004307159 + 0.8536877
> Mf <- function(x)
+ exp(alpha + beta * x) / (1 + exp(alpha + beta * x))
> My <- Mf(x.jiku)
```

```
> points(x.jiku, My, type="l")
```

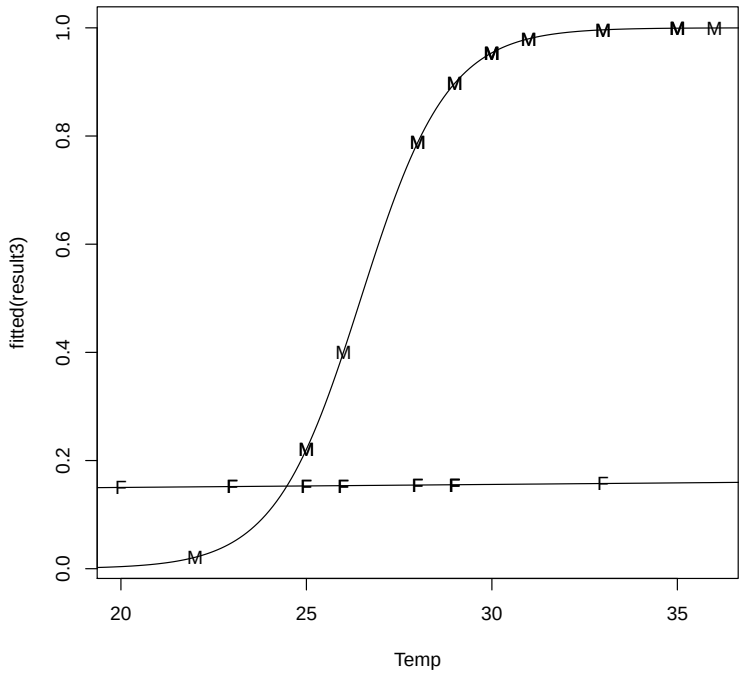


図 4.2　男性と女性の購買行動の予測確率（交互作用モデル）

加法モデルの解釈

加法モデルは交互作用モデルよりも単純である。なぜなら、交互作用モデルでは性別によって定数と係数の両方が影響を受けるが、加法モデルでは定数に影響を受けるだけだからである。

図 4.3 を見ると、加法モデルでは単にロジスティック曲線が x 軸方向へ平行移動しただけであることが分かる。これを見ても、やはり女性は 30 度を超えても購買確率が 0.5 にも満たないので、どちらのモデルが最適であるにせよ、女性のアイスの購買行動は気温に依存していないといえる。

```
> alpha <- -12.31550 # 女性のα
> beta <- 0.3837436 # 女性のβ
> # ロジスティック関数を定義する
> f <- function(x) exp(alpha + beta * x) / (1 + exp(alpha + beta * x))
```

```
> # 横軸に気温、縦軸に予測確率をとってプロット
> plot(Temp, fitted(result2), pch=c("F", "M")[Sex], xlim=c(10, 50))
> xv <- seq(10, 60, 0.1) # x 軸のデータ
> Fyv <- f(xv) # y 軸のデータ
> points(xv, Fyv, type="l") # ロジスティック曲線の当てはめ

> # 以下、男性も同様
> alpha <- -12.31550 + 2.663259
> Myv <- f(xv)
> points(xv, Myv, type="l")
```

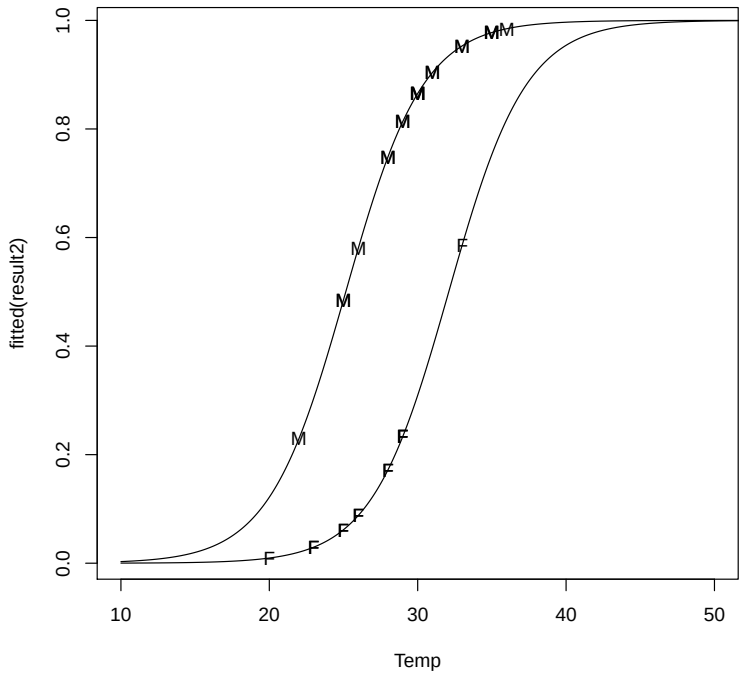


図 4.3　男性と女性の購買行動の予測確率（加法モデル）

4.4　名義・順序ロジットモデル

ロジスティック回帰分析について比率データと 2 値データの取り扱い方を紹介してきた。2 値データとは言い換えれば 2 つの水準をもつカテゴリカル型のデータであると述べたが、3 つ以上の水準をもったカテゴリカル型のデータも取り扱うことができる。それがここで紹介

する名義・順序ロジットモデルである。

名義ロジットモデルは、単純に 2 値ロジットモデルを 3 つ以上の水準をもつカテゴリカル型データへと拡張したものだと考えればよい。一方で順序ロジットモデルは、名義ロジットモデルに、さらに順序関係を考慮した、いわゆる順序尺度のデータを扱うための方法である。

4.4.1　名義ロジットモデル

早速だが表 4.4 を見てみよう。このデータセットは実験参加者の走行時の歩幅を測定して、身長を参考にして $High(175-180cm), Middle(170-174cm), Low(165-169)$ の 3 群にわけてまとめたものである。このデータセットを解析するとき、以下の 2 点について考えることができるだろう。

1. 群によって歩幅の平均値は有意に異なるか
2. 歩幅からどの群に属する確率が高いかを予測できるか

1 つめの問題については 1 要因の分散分析モデルによって明らかにすることができる。つまり、$Group$ は $Stride$ を説明できるか（$Stride = Group$）という問題を取り扱うことになる。これに対して 2 つめの問題は応答変数に $Group$ をおき、説明変数に $Stride$ をおくようなモデル（$Group = Stride$）によって解析することができるだろう。勘のいい読者は気づいただろうが、これは 3 水準のカテゴリカル型変数を応答変数とする**名義ロジットモデル**である。

表 4.4　身長と走行時の歩幅

Group	Stride	Group	Stride	Group	Stride
High	93	Middle	85	Low	66
High	103	Middle	80	Low	67
High	98	Middle	79	Low	79
High	85	Middle	77	Low	69
High	80	Middle	84	Low	70
High	96	Middle	83	Low	71
High	90	Middle	80	Low	68
High	91	Middle	81	Low	65
High	89	Middle	87	Low	70
High	100	Middle	74	Low	66
High	99	Middle	79	Low	78
High	95	Middle	86	Low	77
High	94	Middle	81	Low	60

ところで、もし読者であるあなたが**多変量解析**と呼ばれる諸手法を勉強したことがあるならば、2 つめの問題は判別分析で扱うような問題であると考えたかもしれない。**判別分析**とは既存のデータを基に未知のサンプル（被験体）がどの群に属するかを予測するための分析手法である。例えば、身長と体重によってあるサンプルが男性か女性かを判別できるか、という問題を扱うのが判別分析である。

実をいうと統計学の諸手法は名称（計算理論）が違っても、理屈としては同等なことをやっているものが多かったりする。例えば、**数量化 I 類**は**ダミー変数を用いた回帰分析**と同等であるし、**数量化 II 類**は**ダミー変数を用いた判別分析**である。さらにいえば、ダミー変数を用いた回帰分析というのは、一般線形モデルでいうところの分散分析モデルである。

だから、扱う問題によってはここで紹介する名義ロジットモデルと判別分析による結果が非常に類似したものであることは自然なことである。各手法のアプローチ（理論）が異なるので、解析結果が全く同じにはならないだろうが、実質的には同じことをやっているに過ぎないこともある、ということを覚えておくとよいだろう（似たような手法を色々と見比べてみるのも統計学の醍醐味の 1 つである）。

さて話を元に戻して、表 4.4 のデータセットを名義ロジットモデルで解析してみよう。名義ロジットモデルを扱える関数はいくつかあるが、ここでは VGAM パッケージに含まれている vglm() という関数を利用することにする。なお、VGAM パッケージは標準実装されていないので、ダウンロードしてライブラリから呼び出す必要がある。

```
> install.packages("VGAM") # パッケージのインストール
> library(VGAM) # ライブラリから呼び出す
```

vglm() の使い方は glm() と同じだと考えてよいだろう。ただし、名義ロジットモデルでは誤差が二項分布（binomial）ではなく多項分布（multinomial）に従うと仮定されるので、引数には family=multinomial と指定しなければならないことに注意しよう。また、glm() では応答変数におかれた変数について、最初の水準がベースラインとされたが、vglm() では最後の水準が**ベースライン**とされることを気に留めておこう。

```
> Stride <- matrix(c(
+ 93, 103, 98, 85, 80, 96, 90, 91, 89, 100, 99, 95, 94,
+ 85, 80, 79, 77, 84, 83, 80, 81, 87, 74, 79, 86, 81,
+ 66, 67, 79, 69, 70, 71, 68, 65, 70, 66, 78, 77, 60))
> Group <- rep(1:3, c(13, 13, 13))
> Group <- factor(Group, levels=c(1, 2, 3), labels=c("H", "M", "L"))
```

```
> levels(Group)
[1] "H" "M" "L"

> result <- vglm(Group ~ Stride, multinomial)
> summary(result)

Coefficients:
                  Value Std. Error t value
(Intercept):1 -76.52529   22.20944 -3.4456
(Intercept):2 -38.71092   16.44314 -2.3542
Stride:1        0.94455    0.27442  3.4420
Stride:2        0.50643    0.21153  2.3941
```

名義ロジットモデルでは応答変数の水準数だけ**予測確率**を求めるための係数が出力される。今回の例だと応答変数の水準数は 3 つであるから、ベースラインを除いた 2 つの係数が得られることになる（Intercept が 2 つと Stride の係数が 2 つ得られている）。

そして第 i 水準の予測確率 $\pi_i (i = 1, 2, ..., k)$ は式 4.12 によって求められる（初心者が分かりやすく読めるように、この式は一般的な参考書に載っている数式とは少し異なっている）。

$$\pi_i = \frac{\exp(\alpha_i + \beta_i x)}{1 + \sum_{j=1}^{k-1} \exp(\alpha_j + \beta_j x)} \tag{4.12}$$

ただし、最後の水準の予測確率 π_k は式 4.13 によって求められる（分子が単純に 1 となるのである）。

$$\pi_k = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(\alpha_j + \beta_j x)} \tag{4.13}$$

今回の例で実際に計算してみると分かるだろうが、その前に先ほど得られた係数表をより分かりやすい形で得る方法を紹介しておく。

```
> coef(result, matrix=TRUE)
            log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])
(Intercept)        -76.5252869         -38.710916
Stride               0.9445487           0.506425
```

それではこの係数表をもとに先ほどの数式の意味を解説していこう。まず第 1 水準（$High$）の予測確率 π_1 を求める式は式 4.14 のようになる。

$$\pi_1 = \frac{\exp(-76.525 + 0.944x)}{1 + \exp(-76.525 + 0.944x) + \exp(-38.710 + 0.506x)} \tag{4.14}$$

つぎに第 2 水準（$Middle$）の予測確率 π_2 を求める式は式 4.15 のようになる。

$$\pi_2 = \frac{\exp(-38.710 + 0.506x)}{1 + \exp(-76.525 + 0.944x) + \exp(-38.710 + 0.506x)} \tag{4.15}$$

そして第 3 水準（Low）の予測確率 pi_3 は式 4.16 によって計算することができる。分子が 1 になるのは、第 3 水準をベースラインとしたからである。

$$\pi_3 = \frac{1}{1 + \exp(-76.525 + 0.944x) + \exp(-38.710 + 0.506x)} \tag{4.16}$$

もちろん、実際には `fitted()` を用いれば簡単に予測確率が得られる。ただし、名義ロジットモデルで解析されるようなデータセットでは、そのサンプルサイズ（データ数）が大きい場合が多い。そのため `vglm()` の結果が格納されたオブジェクトを指定すると、恐ろしく膨大なデータ行列が出力されるので注意した方がよい。

もし得られた結果を見たいのであれば、例えば図 4.4 のようなものを作成するなどの工夫をしたほうがよい。これを見れば、H（Heigh）に属する被験者でも確率的には M（Middele）に属するものが 2 人ほどいることがわかる（左上の図における左下に位置する 2 つの H、右上の図における左上に位置する H からそれがわかる）。このことは右下のボックスプロット（箱ひげ図）を見ても、H の下ヒンジが M の中央値に近いことからも、H 群でも歩幅が小さい被験者がいることがわかる。

```
> # 新しいグラフウィンドウを起動
> windows()
> # ウィンドウを 2 行 2 列に分割
> par(mfrow=c(2, 2))

> # 各被験者が H である確率をプロット
> plot(1:length(fitted(result)[,1]), fitted(result)[,1],
+ pch=c("H", "M", "L")[Group])
> # M である確率
> plot(1:length(fitted(result)[,2]), fitted(result)[,2],
+ pch=c("H", "M", "L")[Group])
> # L である確率
```

```
> plot(1:length(fitted(result)[,3]), fitted(result)[,3],
+ pch=c("H", "M", "L")[Group])
> # ボックスプロット
> plot(Group, Stride)
```

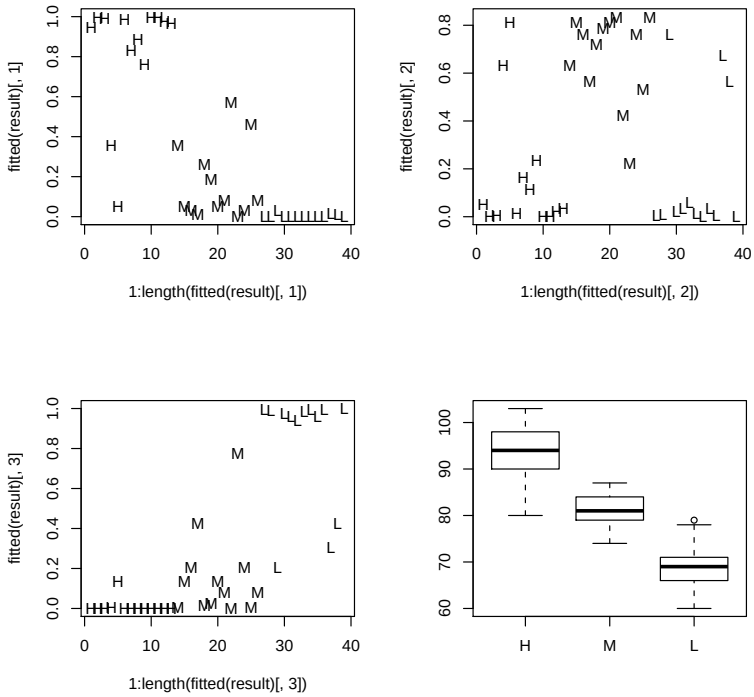


図 4.4　予測確率プロットとボックスプロット

ところで、今回の解析によって 3 本のロジスティック曲線を描く関数が得られた（式 4.14、式 4.15、式 4.16）。これは 2 値ロジットモデルの章で行ったように、1 つ 1 つ自分で関数を定義してもよいが面倒であろう。そのような場合に以下のような方法を覚えておくと便利である。

```
> xv <- seq(60, 110, 0.1)
> length(xv)
[1] 501
> yv <- predict(result, list(Stride=xv), type="response")
> colnames(yv)
[1] "H" "M" "L"
```

```
> length(yv[,1])
[1] 501

> windows()
> par(mfrow=c(2, 2))
> for(i in 1:3){
+ plot(Stride, fitted(result)[,i], pch=c("H", "M", "L")[Group])
+ points(xv, yv[,i], type="l")
+ }
> plot(Group, Stride)
```

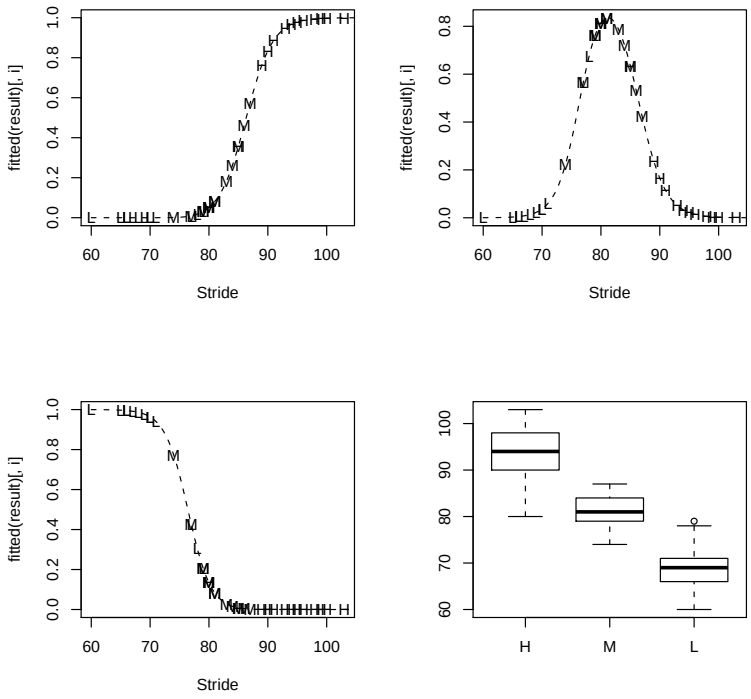


図 4.5　水準ごとのロジスティック曲線

コツは predict() を上手く使うことで、上の実行例では引数に list(Stride=xv) と指定している。簡単にいってしまうと、これは「Stride の変わりに xv の値を使いなさい」ということを命令しているのである。これを利用すると、例えば歩幅が 73cm である被験者が 3 つの水準 H, M, L それぞれに属する確率を得ることができる。

```
> predict(result, list(Stride=c(73)), type="response")
             H         M        L
1 0.0004372108 0.1490078 0.850555
```

さらにちょっとした工夫を加えれば、いろいろと考察するための材料が得られる。以下の出力から、歩幅が 75cm くらいを境に L に属するか否かの確率が大きく変化することが分かる。また、70cm のときに 0.96 で、80cm のときに 0.13 になっていることから、L に属する確率を描くロジスティック曲線は 70 から 80 までの間で S 字カーブを描くのだろうと予想できる。実際に図 4.5 の左下の曲線を見てもそのようになっていることがわかり、こういったちょっとした作業はグラフを描くときにも（軸の設定などで）役に立ったりする利点もある。

```
> a <- predict(result, list(Stride=c(70:80)), type="response")
> cbind(a, 70:80)
               H          M         L
1  2.910702e-05 0.03692666 0.9630442 70
2  7.307072e-05 0.05981511 0.9401118 71
3  1.807626e-04 0.09547760 0.9043416 72
4  4.372108e-04 0.14900779 0.8505550 73
5  1.023131e-03 0.22499583 0.7739810 74
6  2.288027e-03 0.32466078 0.6730512 75
7  4.832226e-03 0.44242655 0.5527412 76
8  9.564174e-03 0.56502376 0.4254121 77
9  1.772564e-02 0.67568854 0.3065858 78
10 3.093868e-02 0.76097697 0.2080844 79
11 5.131893e-02 0.81446547 0.1342156 80
```

4.4.2 順序ロジットモデル

応答変数に用いるカテゴリカル型のデータに順序関係があるならば、その情報を取り入れた解析を行った方が良いこともある（結果としてどちらでも変わりがないこともある）。先ほどの名義ロジットモデルでの解析では、変数 $Group$ は $High, Middle, Low$ といった 3 つの水準をもっていた。しかし、これらの水準間に順序関係はないものとして扱っていた。そこで今回は $Low < Middle < High$ という順序関係の情報を付加して解析してみよう。

まず解析に先立って、順序尺度のデータを用意する方法を紹介する。名義尺度のデータを名義ロジットモデルで扱った Group とし、新たに順序関係の情報を付加した Group2 を用意する。変数に順序情報を付加するには関数 ordered() を使うか、あるいは factor() の引数に ordered=TRUE と書き加えればよい。

```
> Stride <- matrix(c(
+ 93, 103, 98, 85, 80, 96, 90, 91, 89, 100, 99, 95, 94,
+ 85, 80, 79, 77, 84, 83, 80, 81, 87, 74, 79, 86, 81,
+ 66, 67, 79, 69, 70, 71, 68, 65, 70, 66, 78, 77, 60))
> Group <- rep(1:3, c(13, 13, 13))
> Group <- factor(Group, levels=c(1, 2, 3), labels=c("H", "M", "L"))
> Group
Levels: H M L

> Stride2 <- matrix(c(
+ 66, 67, 79, 69, 70, 71, 68, 65, 70, 66, 78, 77, 60,
+ 85, 80, 79, 77, 84, 83, 80, 81, 87, 74, 79, 86, 81,
+ 93, 103, 98, 85, 80, 96, 90, 91, 89, 100, 99, 95, 94))
> Group2 <- rep(1:3, c(13, 13, 13))
> Group2 <- factor(Group2, levels=c(1, 2, 3),
+ labels=c("L", "M", "H"), ordered=TRUE)
> # Group2 <- orderd(Group2, levels=1:3, labels=c("L", "M", "H")
> Group2
Levels: L < M < H
```

累積ロジットモデル

累積ロジットモデルとして解析するためには、vglm() で family=cumulative と指定する。なお、parallel を省略した場合は FALSE となる（デフォルトでは parallel=FALSE になっている）。

```
> clm <- vglm(Group2 ~ Stride2, cumulative(parallel=FALSE))
> summary(clm)
```

```
Coefficients:
                 Value Std. Error t value
(Intercept):1 39.31315   16.00422  2.4564
(Intercept):2 39.33490   14.33978  2.7431
Stride2:1     -0.51450    0.20555 -2.5030
Stride2:2     -0.45518    0.16760 -2.7160

Residual Deviance: 28.76714 on 74 degrees of freedom

> coef(clm, matrix=TRUE)
            logit(P[Y<=1]) logit(P[Y<=2])
(Intercept)     39.3131507     39.3349003
Stride2         -0.5144983     -0.4551832
```

比例オッズモデル

比例オッズモデルとして解析するためには、parallel=TRUE と指定する。

```
> pom <- vglm(Group2 ~ Stride2, cumulative(paralle=TRUE))
> summary(pom)

Coefficients:
                 Value Std. Error t value
(Intercept):1 36.77217   10.12824  3.6307
(Intercept):2 41.60830   11.08910  3.7522
Stride2       -0.48172    0.12972 -3.7135

Residual Deviance: 28.81773 on 75 degrees of freedom

> coef(pom, matrix=TRUE)
            logit(P[Y<=1]) logit(P[Y<=2])
(Intercept)     36.7721700     41.6082965
Stride2         -0.4817237     -0.4817237
```

分割表への適用例

最後に分割表の解析にも利用することができるという例を示しておこう。例えば、表 4.5 のような分割表データがあった場合、名義ロジットモデルや順序ロジットモデルを用いて各セルの確率を求めるといったアプローチができる。

表 4.5　1 週間に自炊する回数

	ほぼ毎日	週に 2,3 回	ほとんどない
男性	45	121	33
女性	40	110	23

ここでは名義ロジットモデル、累積ロジットモデル、比例オッズモデルの 3 つの方法を用いて解析してみる。表表 4.5 のようなデータセットを、あらかじめテキストファイルなどとして保存しておき、`read.table()` などで読み込んだほうが手間が省けるだろう。

```
> # 解析に使用するデータセット
> dat
  ほぼ毎日 週に 2.3 回 ほとんどない     性別
1       45       121           33   male
2       40       110           54 female

> # 名義ロジットモデル
> tbl.mlm <- vglm(cbind(dat[,1], dat[,2], dat[,3]) ~ dat[,4],
+ multinomial)
> # 累積ロジットモデル
> tbl.clm <- vglm(cbind(dat[,1], dat[,2], dat[,3]) ~ dat[,4],
+ cumulative)
> # 比例オッズモデル
> tbl.pom <- vglm(cbind(dat[,1], dat[,2], dat[,3]) ~ dat[,4],
+ cumulative(parallel=TRUE))

> coef(tbl.mlm, matrix=TRUE) # 名義ロジットの係数表
             log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])
(Intercept)          -0.3001046          0.7114963
dat[, 4]male          0.6102595          0.5877867
```

```
> coef(tbl.clm, matrix=TRUE) # 累積ロジットの係数表
             logit(P[Y<=1]) logit(P[Y<=2])
(Intercept)      -1.4109870       1.021651
dat[, 4]male      0.1806969       0.593829
> coef(tbl.pom, matrix=TRUE) # 比例オッズの係数表
             logit(P[Y<=1]) logit(P[Y<=2])
(Intercept)      -1.5255034      1.1051962
dat[, 4]male      0.3870449      0.3870449

> fitted(tbl.mlm) # 名義ロジットによる予測確率
        mu1       mu2       mu3
1 0.2261307 0.6080402 0.1658291
2 0.1960784 0.5392157 0.2647059
> fitted(tbl.clm) # 累積ロジットによる予測確率
        mu1       mu2       mu3
1 0.2261307 0.6080402 0.1658291
2 0.1960784 0.5392157 0.2647059
> fitted(tbl.pom) # 比例オッズによる予測確率
        mu1       mu2       mu3
1 0.2426035 0.5738109 0.1835856
2 0.1786525 0.5725799 0.2487676
```

名義ロジットモデルと累積ロジットモデルによる予測確率は同じであるが、比例オッズモデルによる予測確率は少し異なっているのが分かる。

順序ロジットモデルには、ここで紹介した 2 つのモデル以外にも隣接カテゴリーロジットモデルや連続比ロジットモデルといったモデルがある。関心のある読者は文献 [2] や文献 [4] を参照するとよいだろう。

第 5 章

さらに高度なモデル解析

今までに扱ってきたモデルは全て独立なデータに対して適用されるべきモデルであった。原則として統計モデルは得られている変数同士が独立であることが前提とされているが、心理学の分野ではしばしば対応ありのデータが得られる。対応ありのデータとは経時測定データ、あるいは繰返測定データと呼ばれている、ということはすでに述べた（本書の 2 **章** 4 **節** **反復測定と繰返測定**を参照のこと）。

ここでは**経時測定データ**の解析法として、一般線形モデルの分散分析モデルとして扱う方法と、**一般線形混合モデル**と呼ばれる新たなモデルを紹介する。さらに**多変量解析モデル**のなかでも、特に心理学の分野で多用される**因子分析モデル**と**主成分分析モデル**、判別分析モデルの 2 つを紹介する。さらに本章では、多変量解析の結果だけでなく、その結果を 2 時データとしてさらに高度に扱う方法も示す。

拙者は以下に紹介するこの 3 つのモデルはかなり高度なモデル解析として位置づけている。したがって、ここまで扱ってきた一般線形モデル（分散分析、回帰分析、共分散分析）や一般化線形モデル（ポアソン回帰分析、ロジスティック回帰分析）の章を熟読してから読み進めることを薦める。何事も「とりあえず練習でやってみる」ということは重要だが、あまり焦りすぎて一度に多くのものを習得しようとしても、そうそうは身につくものではないからである。

実際、一般線形モデルを十分に理解しているならば、心理学論文をかなり高度に読み解くことができる（一般化線形モデルを用いている心理学論文の数は限られているからである）。だから、まずは単純なモデルだけでもよいから、モデル解析という考え方をきちんと身につけるべきなのである。もしも本書に書いてある内容を理解できているならば、心理学論文を読んで、その論文でどのようなモデルが扱われているのかをモデル式で表現できるはずである（逆にいえば、他者の論文を読んでモデル式を書けないようなら、まだ訓練が足りないと自覚すべきである）。

5.1　経時測定データの解析法

5.1.1　データの差をとる方法

経時測定データを扱う方法の 1 つとして、最初に**データの差をとる方法**を紹介する。例えば、表 5.1 のようなデータセットについて、この方法を適用してみよう。これは高血圧と判断された 18 人を対象に、降圧剤を 5mg、10mg、20mg とそれぞれ 6 人ずつの 3 グループに別けて、投与してから 5 週間の血圧の変化を測定したものである。

表 5.1　降圧剤の投与量と血圧の変化（経時測定データ）

1WEEK	2WEEK	3WEEK	4WEEK	5WEEK	DOSE
150	145	130	133	130	5mg
155	144	134	130	125	5mg
149	149	133	130	125	5mg
160	147	129	128	123	5mg
153	144	140	135	130	5mg
155	146	136	133	127	5mg
157	143	147	129	110	10mg
160	148	140	130	112	10mg
160	156	144	124	120	10mg
155	150	150	120	119	10mg
149	149	146	124	120	10mg
148	150	135	119	120	10mg
151	153	139	110	90	20mg
160	149	147	112	99	20mg
157	155	144	120	100	20mg
154	149	138	123	98	20mg
151	144	140	128	101	20mg
150	150	136	122	103	20mg

もしも分析者が 1 週間目の血圧と 5 週間目の血圧の変化に関心があるならば、5 週目の血圧と 1 週間目の血圧の差をとってやればよい。例えば 1 人目の被験者ならば、$130-150=-20$ となるので、この -20 という差のデータを応答変数として用いればよいわけである。ただし、この方法では 2 週間目から 4 週間目のデータを無駄にしていることになる。したがって、この方法が有用なのは、いわゆる**事前・事後データ**を解析するときである（事前・事後データは古典的に**対応のある** t **検定**が用いられる）。

とりあえず有用であるかどうかは別問題として、この方法を用いた解析例を示すことにする。ここでは表 5.1 のデータが dat という変数に格納されていることとする。なお、R では変数名の先頭に数字が使えないので、1WEEK という変数名は使えない。関数 read.table() で読み込んだ場合には、自動的に変数の先頭に X が付け加えられて X1WEEK といったようになる。

```
> dat
   X1WEEK X2WEEK X3WEEK X4WEEK X5WEEK DOSE
1     150    145    130    133    130  5mg
2     155    144    134    130    125  5mg
3     149    149    133    130    125  5mg
(途中省略)

> dif <- dat$X5WEEK - dat$X1WEEK # データの差をとる
> dif
 [1] -20 -30 -24 -37 -23 -28 -47 -48 -40
 [10] -36 -29 -28 -61 -61 -57 -56 -50 -47
> dat$DOSE
 [1] 5mg  5mg  5mg  5mg  5mg  5mg  10mg 10mg 10mg
 [10 10mg 10mg 10mg 20mg 20mg 20mg 20mg 20mg 20mg
Levels: 10mg 20mg 5mg

> result <- lm(dif ~ dat$DOSE)
> summary(aov(result))
            Df  Sum Sq Mean Sq F value    Pr(>F)
dat$DOSE     2 2448.44 1224.22  25.528 1.484e-05 ***
Residuals   15  719.33   47.96

> summary(result)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -38.000      2.827 -13.441 9.04e-10 ***
dat$DOSE20mg  -17.333      3.998  -4.335 0.000588 ***
dat$DOSE5mg    11.000      3.998   2.751 0.014848 *
```

分散分析表によれば、投与量（$DOSE$）によって血圧の降下は有意に認められるということが主張できる（$p = 0.0000$）。そして係数表から、この場合は 10mg という水準をベースラインとしているので、もっとも降圧の効果が認められるのは 20mg であることがわかる（dat$DOSE20mg: $p = 0.0006$）。5mg の場合では 10mg に比べて降圧効果が小さいことも見てとれる（dat$DOSE5mg: $p = 0.0015$）。

5.1.2　多変量分散分析モデル

経時測定データを一般線形モデルの枠組みで解析するもう 1 つの方法が**多変量分散分析モデル**として解析することである。多変量分散分析とは、応答変数が多変量である場合の分散分析である。今回の解析で扱うモデル式は以下のように書くことができる。

$$1W, 2W, 3W, 4W, 5W = DOSE \tag{5.1}$$

R でこのモデルを解析するためには manova() という関数を用いる。使い方は lm() とほぼ一緒で、応答変数を行列で指定するという点のみが異なる。以下にその使用例を示す。

```
> res.dat <- cbind(dat$X1WEEK, dat$X2WEEK, dat$X3WEEK,
+ dat$X4WEEK, dat$X5WEEK)
> res.dat
      [,1] [,2] [,3] [,4] [,5]
 [1,]  150  145  130  133  130
 [2,]  155  144  134  130  125
 [3,]  149  149  133  130  125
（途中省略）

> manova.result <- manova(res.dat ~ dat$DOSE)
> summary(manova.result, test="Wilks") # 多変量分散分析表
          Df  Wilks approx F num Df den Df    Pr(>F)
dat$DOSE   2 0.0360   9.3907     10     22 7.334e-06 ***
Residuals 15

> summary(aov(manova.result)) # 週ごとの分散分析表
 Response 1 :
            Df  Sum Sq Mean Sq F value Pr(>F)
dat$DOSE     2   4.778   2.389  0.1207 0.8872
Residuals   15 297.000  19.800
```

```
 Response 2 :
            Df  Sum Sq Mean Sq F value Pr(>F)
dat$DOSE     2  60.111  30.056  2.5304  0.113
Residuals   15 178.167  11.878

 Response 3 :
            Df  Sum Sq Mean Sq F value   Pr(>F)
dat$DOSE     2 316.000 158.000  7.6452 0.005139 **
Residuals   15 310.000  20.667

 Response 4 :
            Df Sum Sq Mean Sq F value   Pr(>F)
dat$DOSE     2 460.33  230.17  9.2892 0.002373 **
Residuals   15 371.67   24.78

 Response 5 :
            Df  Sum Sq Mean Sq F value    Pr(>F)
dat$DOSE     2 2452.33 1226.17  74.263 1.655e-08 ***
Residuals   15  247.67   16.51
```

さて上のアウトプットの多変量分散分析表を見ると、投与量によって降圧効果が有意に認められるということが分かる（$p = 0.0000$）。さらに詳しく週ごとの分散分析表を見ていくと、投与量の水準間で差が認められるのは 3 週目（`Response 3 :`）からであることがわかる。つまり、2 週目までは投与量の違いによる変化は認められないが、3 週目以降に大きく投与量の違いが現れるということがわかる（試しに $DOSE$ の平方和: `dat$DOSE` の `Sum Sq` をプロットしてみるとよい）。

5.1.3 一般線形混合モデル

先に述べたデータの差をとる方法と多変量分散分析による解析は一般線形モデルの枠組みで解析できる方法の例である。これらとは別に**一般線形混合モデル**による解析法が存在する。これは簡単にいってしまえば、被験者ごとにモデルを当てはめるというようなやり方である。換言すると、被験者ごとに予測値を求めるという作業を行うわけである。

そこで一般線形混合モデルには、被験者によって切片のみがランダムに変化するという仮定をする**ランダム切片モデル**と、被験者によって切片と傾きの両方がランダムに変化すると

仮定する**ランダム切片・傾きモデル**という 2 つのモデルが考えられる。

ランダム切片モデル

表 5.1 のデータテーブルはごく一般的な形式としてまとめられている。しかしモデル解析を行うという立場では、このようなまとめ方はあまり好ましいものではない。そこで表 5.1 のデータテーブルを表 5.2 のようなデータフレームの形式に直してみよう。SBJ は被験者（Subjects)、$WEEK$ は経過週、$DOSE$ は投与量、$BLPR$ は血圧（Blood Pressure）である。

表 5.2　降圧剤の投与量と血圧の変化（データフレーム）

SBJ	WEEK	DOSE	BLPR
1	1	5mg	150
1	2	5mg	145
1	3	5mg	130
1	4	5mg	133
1	5	5mg	130
2	1	5mg	155
2	2	5mg	144
2	3	5mg	134
2	4	5mg	130
2	5	5mg	125
	（途中省略）		
18	1	20mg	150
18	2	20mg	150
18	3	20mg	136
18	4	20mg	122
18	5	20mg	103

もしも完全に独立なデータであるならば、式 5.2 のような 2 要因の分散分析モデルを解析することになる（つまり α, β_1, β_{2j} というパラメータを推定することになる)。α は定数項（切片)、β_1 は $WEEK$ の第 i 水準の係数（傾き)、β_{2j} は $DOSE$ の第 j 水準の係数（傾き）である。

$$BLPR = \alpha + \beta_{1i} WEEK + \beta_{2j} DOSE \tag{5.2}$$

これをランダム切片モデルとして解析する場合、新たに SBJ という変数をモデルに投入することになる。この SBJ は**ランダム要因** (**変量効果**) と呼ばれるもので、SBJ は平均が 0、分散が σ^2_{SBJ} の正規分布に従うと仮定される。このモデルは式 5.3 のように書ける。

$$BLPR = \alpha + \beta_{1i} WEEK + \beta_{2j} DOSE + SBJ \tag{5.3}$$

これを次式 5.5 のように書き直してみるとより分かりやすいだろう。要するに、これは切片の値がランダムに変化することを考慮したモデルなのである（少し違う言い方をすると個体間変動を考慮しているのである)。つまり SBJ はランダムに変化しうる要因だと考えているわけである。換言すれば、被験者はランダムに選ばれたものであるから、異なる被験者が選ばれるたびに異なる値をとるだろうというわけである。

$$BLPR = (\alpha + SBJ) + \beta_{1i} WEEK + \beta_{2j} DOSE \tag{5.4}$$

とりあえず、R による実行例を示して、それを見ながら詳しく説明していくことにする。R で一般線形混合モデルを扱うには `lme()` という関数を用いるが、これは標準実装されていないので、`install.packages("nlme")` とコマンドした後、`library(nlme)` としてライブラリから呼び出さなければ使えない。また、`lme()` の引数で、ランダム要因である SBJ を `random= 1|SBJ` として指定しなければならないことに注意しよう。

```
> WEEK <- rep(1:5, 18)
> WEEK <- factor(WEEK, levels=1:5,
+ labels=c("1W", "2W", "3W", "4W", "5W"))
> DOSE <- rep(1:3, c(30, 30, 30))
> DOSE <- factor(DOSE, levels=1:3, labels=c("5mg", "10mg", "20mg"))
> SBJ <- rep(1:18, c(rep(5, 18)))
> SBJ <- as.factor(SBJ)
> dat <- matrix(c(
+ 150, 145, 130, 133, 130,
+ 155, 144, 134, 130, 125,
+ 149, 149, 133, 130, 125,
+ 160, 147, 129, 128, 123,
+ 153, 144, 140, 135, 130,
+ 155, 146, 136, 133, 127,
+ 157, 143, 147, 129, 110,
+ 160, 148, 140, 130, 112,
+ 160, 156, 144, 124, 120,
+ 155, 150, 150, 120, 119,
+ 149, 149, 146, 124, 120,
+ 148, 150, 135, 119, 120,
```

```
+ 151, 153, 139, 110, 90,
+ 160, 149, 147, 112, 99,
+ 157, 155, 144, 120, 100,
+ 154, 149, 138, 123, 98,
+ 151, 144, 140, 128, 101,
+ 150, 150, 136, 122, 103), ncol=5, byrow=TRUE)
> BLPR <- c(t(dat))

> library(nlme)
> model1 <- lme(BLPR ~ WEEK + DOSE, random=~1|SBJ, method="ML")
> summary(model1)

Random effects:
 Formula: ~1 | SBJ
         (Intercept) Residual
StdDev: 0.0002593324 6.722755

Fixed effects: BLPR ~ WEEK + DOSE
                Value Std.Error DF   t-value p-value
(Intercept) 156.21111  1.952348 68  80.01192  0.0000
WEEK2W       -5.72222  2.333502 68  -2.45220  0.0168
WEEK3W      -14.77778  2.333502 68  -6.33287  0.0000
WEEK4W      -29.11111  2.333502 68 -12.47529  0.0000
WEEK5W      -40.11111  2.333502 68 -17.18923  0.0000
DOSE10mg     -0.46667  1.807523 15  -0.25818  0.7998
DOSE20mg     -5.83333  1.807523 15  -3.22725  0.0056

> summary(aov(model1))
            Df  Sum Sq Mean Sq  F value    Pr(>F)
WEEK         4 19754.4  4938.6 100.7734 < 2.2e-16 ***
DOSE         2   630.5   315.2   6.4324  0.002529 **
Residuals   83  4067.6    49.0
```

ここで Fixed effects:　とある部分は通常の分散分析モデルの解析結果と同じ係数表である（もちろん lm() を使って解析した場合とで数値は異なるが）。問題は Random effects:

という部分である。これは分散成分と呼ばれるもので、Intercept の値が SBJ の分散であり、Residual が誤差分散である。

この 0.0002593324 という値は被験者間のバラつきを表しているのである。この場合、バラつきはほとんど 0 なので、被験者ごとに当てはめられる理論直線（折れ線）はほぼ重なるかたちになる。仮にこの値が大きければ、当てはめられる理論直線は重なって見えないだろう（平行な何本もの理論直線が引かれることになる）。

これは実際に予測値を図で表してみるとよく分かるだろう（図 5.1）。観測値（原データ）と予測値を群別にプロットするには matplot() を利用すると便利である。以下にその例を示す。

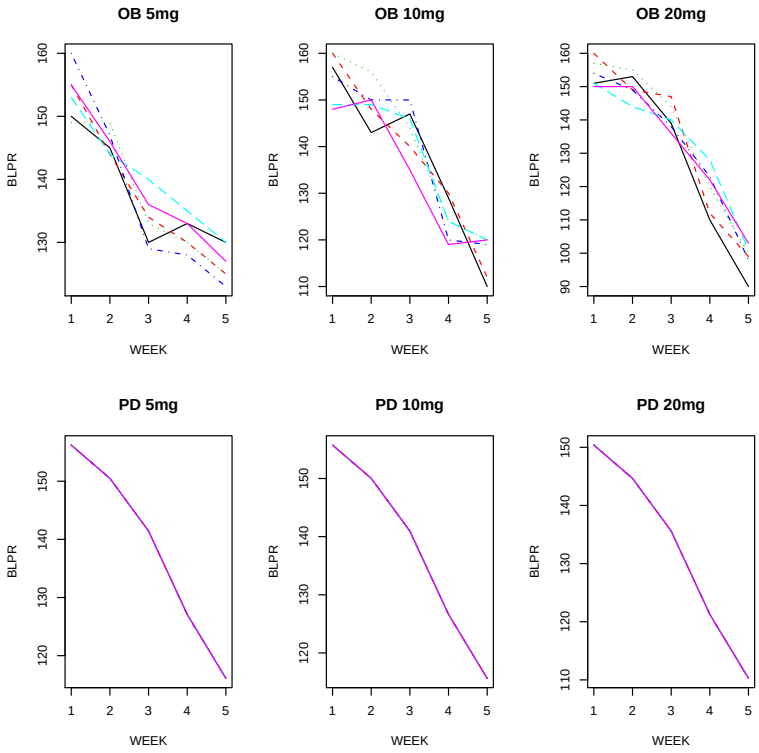


図 5.1　ランダム切片モデルの当てはめ

```
> # 予測値を res1 に代入
> res1 <- matrix(predict(model1), ncol=5, byrow=TRUE)

> # 観測値のプロット
> matplot(1:5, t(dat[1:6, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="OB 5mg")
> matplot(1:5, t(dat[7:12, 1:5]),
```

```
+ type="l", xlab="WEEK", ylab="BLPR", main="OB 10mg")
> matplot(1:5, t(dat[13:18, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="OB 20mg")

> # 予測値のプロット
> matplot(1:5, t(res1[1:6, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="PD 5mg")
> matplot(1:5, t(res1[7:12, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="PD 10mg")
> matplot(1:5, t(res1[13:18, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="PD 20mg")
```

ランダム切片・傾きモデル

先ほどの解析をランダム切片・傾きモデルとして改めて解析してみよう。このモデルでは、傾きもランダムに変動するという仮定をおくことになり、モデル式は以下のようになる。

$$BLPR = (\alpha + SBJ) + (\beta_{1i} + \lambda)WEEK + \beta_{2j}DOSE \tag{5.5}$$

以下に R による実行例を示す。

```
> model2 <- lme(BLPR ~ WEEK + DOSE, random=~WEEK|SBJ, method="ML")
> summary(model2)

Random effects:
 Formula: ~WEEK | SBJ
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept)  3.701966 (Intr) WEEK2W WEEK3W WEEK4W
WEEK2W       4.119442 -0.543
WEEK3W       5.275545 -0.176  0.409
WEEK4W       7.384751 -0.669 -0.180 -0.278
WEEK5W      12.904784 -0.686 -0.037 -0.238  0.810
Residual     2.174068
```

```
Fixed effects: BLPR ~ WEEK + DOSE
                Value Std.Error DF   t-value p-value
(Intercept) 155.10115  1.182628 68 131.14960  0.0000
WEEK2W       -5.72222  1.261643 68  -4.53553  0.0000
WEEK3W      -14.77778  1.498685 68  -9.86050  0.0000
WEEK4W      -29.11111  1.963334 68 -14.82739  0.0000
WEEK5W      -40.11111  3.256010 68 -12.31910  0.0000
DOSE10mg      0.17843  0.929999 15   0.19187  0.8504
DOSE20mg     -3.14854  0.929999 15  -3.38553  0.0041
```

さて Random effects: における Intercept は先ほどと同じく、切片のバラつきを表している。今度は 3.70 と先ほどの値（0.00）よりも大きいので、当てはめられる理論直線は重ならないようになる。

また、ここでは新たに $WEEK$ をランダム変数として組み込んだのでそれについての結果も出力されている。例えば、WEEK2W = 4.11 という値が得られているが、これは 2W の水準における被験者間のバラつきを表しているのである。

週の経過に伴って分散が大きくなっていることから、週が経過するほど各被験者に対する降圧剤の効果にバラつきが生じているということがわかる。この切片のバラつきが大きいほど、当てはめられる理論直線（折れ線）の形状は被験者によって大きく異なったものとなる。仮にバラつきが 0 に近ければ、被験者ごとの理論直線の形状はほぼ同じものとなるだろう（つまりランダム切片モデルと同じような結果になるわけである）。

```
> res2 <- matrix(predict(model2), ncol=5, byrow=TRUE)

> matplot(1:5, t(dat[1:6, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="OB 5mg")
> matplot(1:5, t(dat[7:12, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="OB 10mg")
> matplot(1:5, t(dat[13:18, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="OB 20mg")

> matplot(1:5, t(res2[1:6, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="PD 5mg")
> matplot(1:5, t(res2[7:12, 1:5]),
```

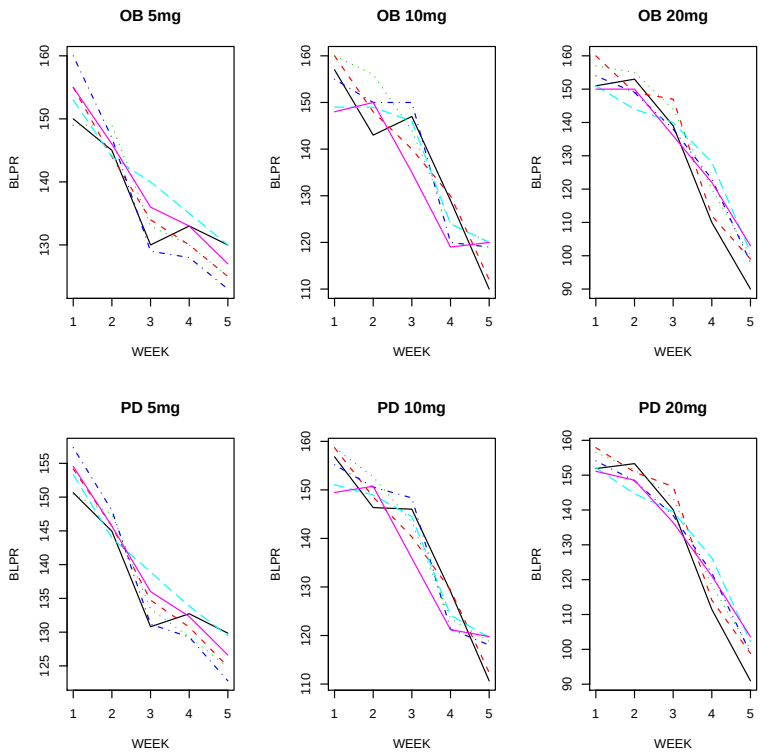


図 5.2　ランダム切片・傾きモデルの当てはめ

```
+ type="l", xlab="WEEK", ylab="BLPR", main="PD 10mg")
> matplot(1:5, t(res2[13:18, 1:5]),
+ type="l", xlab="WEEK", ylab="BLPR", main="PD 20mg")
```

モデルの比較

ここではランダム切片モデルとランダム切片・傾きモデルの 2 つを扱ったが、どちらのモデルを当てはめた方がより適切なのであろうか。この疑問に対する答えを得るためには anova() を用いて、2 つのモデルの比較を行う必要がある。

これは「加えられた項は有用ではない」という帰無仮説について検定されるものである。以下の結果では、$p = 0.0001$ となっているので、帰無仮説を棄却し、「加えられた項は有用である」という結論が得られる。つまり、$WEEK$ を新たにランダム要因として組み込んだ、ランダム切片・傾きモデルの方がより適切なモデルであることを意味している。

```
> anova(model1, model2)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
```

```
model1     1  9 616.3986 638.8969 -299.1993
model2     2 23 588.3309 645.8265 -271.1655 1 vs 2 56.06766  <.0001
```

混合モデルを扱う上での注意点

一般にランダム要因をモデルに組み込むと結果の解釈が難しくなる。ここでは最も単純な例をあげたが、例えば、$WEEK$ と $DOSE$ の交互作用項を組み込むモデルや $WEEK$ の 2 次項を組み込むモデルも考えられる。もしも観測値のプロットを見て、曲線関係がありそうなら $WEEK$ の 2 次項を組み込む必要があるだろう（ただし、$WEEK$ が連続型であるとした場合)。また、ここではカテゴリカル型の変数を 2 つ用いたが、カテゴリカル型と連続型を混在させることも一般線形モデルと同じように可能である。

しかし、いかに複雑でより当てはまりのよいモデルを解析できたとしても、それを分析者自身が適切に解釈できないようでは本末転倒である。ここで気に留めておくべき教訓は「自分が適切に解釈できる統計モデルにはどのようなモデルがあるだろうか」と自問自答するべき、ということである。そして、この教訓は必ず実験計画の段階で考えなければならない。とりあえずデータをとってはみたが、後になってやはり難しいモデル解析になってしまうので困った、となっても後の祭りだからである。

実行されるべきデータ解析の方法が、自分が自信を持って扱えるものでなければ、実験の計画を練りなおしてより単純なモデル解析ができないかを考えるべきである。あるいは高度な統計モデルを理解できるように勉強しなければならない。それをせずに多くの統計モデルを使いこなそうななど、土台無理な話である。だから、まずは単純なモデルでもよいから、それを使えるような実験計画を立てるように心がけてほしい（残念ながら心理学の分野では、データ解析の方法をデータをとった後に考える者が非常に多い)。

5.2 因子分析モデル

ここまでは現実に得られた測定値、すなわち**観測変数**を用いた解析法を紹介してきた。しかし、統計学には測定値として得られていない**潜在変数**と呼ばれる変数を扱う手法が存在する。それがこれより紹介する因子分析や主成分分析である。

5.2.1 因子分析が扱う問題

心理学の分野において、因子分析は質問紙の項目分類のために用いられることが多い。例えば、ある大学生を対象に将来的に子供を持ったとき、子供を持つことに対してメリットを感じるか、それともデメリットを感じるかを調べたいとしよう。

この場合、子供に対するメリットとして「家族がにぎやかになる」や「夫婦の愛情がより

深まる」というような質問項目があげられるかもしれない。一方で子供に対するデメリットを表す質問項目では「子育てにかかる費用が多い」とか、あるいは女性であれば「夫が子育てに参加してくれるかどうか不安である」という項目が考えられるだろう。

このようにメリットに関する質問項目群とデメリットに対する質問項目群を用意したとする。そこで因子分析をすることによって、どの質問項目が子供に対するメリットを表す項目なのか、それともデメリットを表す項目なのかを明らかにするわけである。

5.2.2 因子分析のモデル式

まずは**因子分析モデル**とはどういったモデルなのかを考えてみよう。結論からいって、因子分析モデルとは、ある観測変数を応答変数とし、ある潜在変数を説明変数とした回帰分析モデルなのである。

今回のケースでは子供を持つことのメリットに関する項目、およびデメリットに関する項目すべてが観測変数である。これに対して潜在変数は「メリット」と「デメリット」である。因子分析モデルでは、実際には観測できないが、モデル作成者が「在る」と想定する変数を導入する。これを実際に観測された変数と区別するために、潜在変数と称しているわけである。

ここでは「メリット」とか「デメリット」という変数は存在していない (観測されていない) が、分析者が「この項目群は総称してメリット (あるいはデメリット) という変数として扱えるだろう」と仮定している。

さて、すでに述べたように、因子分析モデルとは観測変数を応答変数、潜在変数（因子）を説明変数とした回帰分析モデルである。したがって、因子分析モデルはいくつかの回帰分析モデルの集合であるといえる。仮に 2 つの観測変数 X_1, X_2 が第 1 因子 F_1 に含まれ、別のもう 2 つの観測変数 X_3, X_4 が第 2 因子 F_2 に含まれるような、いわゆる 2 因子構造をもった因子分析モデルは次のように書ける。

$$
\begin{aligned}
X_1 &= \beta_1 \times F_1 + e_1 \\
X_2 &= \beta_2 \times F_1 + e_2 \\
X_3 &= \beta_3 \times F_2 + e_3 \\
X_4 &= \beta_4 \times F_2 + 3_4
\end{aligned}
\tag{5.6}
$$

これらは一般線形モデルと一般化線形モデルのところで嫌というほど見てきたものであろう。ただし、因子分析を初めとする多変量解析のほとんどは**標準化解**が推定されることになるので、モデル式に定数項は書かない。

少し補足しておくと、標準化解とは全ての元のデータを平均 0、標準偏差 1 となるように変換（標準化、基準化）したデータに対して分析を行った場合の係数のことである。因子分析においては特に**因子負荷量**というのが慣わしである。

標準化解については一般線形モデル、一般化線形モデルのところで特に解説しなかった。

この標準化解を用いる必要があるのは、複数の説明変数を比較してどの変数が相対的に応答変数へ影響を与えているかということを判断するときである。したがって、重回帰モデルの解析のときに用いるだけで、他の場合ではこのような作業をすることはないだろう。

例えば、重さ（kg）と長さ（cm）はそれぞれ単位が異なるので、直接的に両者を比較しても意味がない。そこでデータを標準化して回帰分析を行うことによって、標準化偏回帰係数が得られるのでそれを比較するわけである。しかし、実際には変換公式があるので、わざわざデータを標準化して分析をやり直す必要はない。

回帰方程式 $Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$ において、標準化偏回帰係数 $\beta_i\prime$ は次式によって計算される。

$$\beta_i\prime = \beta_i \times \frac{sd(X_i)}{sd(Y)} \tag{5.7}$$

いうまでもないが、標準化偏回帰係数は直接に予測には役立たない。定数項を 0 とした標準化偏回帰係数を採用したモデル式にデータを代入して予測値を計算しても、それは実測値とはかけ離れた数値になる（当たり前だが)。標準化偏回帰係数はあくまでも変数同士の寄与度の比較をするためのものであって、予測値を求める場合は定数項を含む偏回帰係数によるモデル式を採用する必要がある。

話を元に戻そう。上に挙げた因子分析モデルには 4 つの回帰モデルがあり、β_1 から β_4 までのパラメータ（因子負荷量）を求めることになる。そのためには R に標準実装されている `factanal()` という関数を用いればよい。

ここでは子供を持つことに対するメリットとしてあげられる 4 つの質問項目と、デメリットとして考えられる 4 つの質問項目、計 8 個の項目を観測変数として得られたデータセット `kodomo.txt`[*1]を用いる。このデータセットにはこの 8 個の変数とは別に、子供が欲しいという願望の程度を表した変数 *desire* が含まれている。この *desire* の得点が高いほど、将来的に子供を欲しがっているということができる。

また、これらの質問項目は、あらかじめ 2 因子構造を仮定して用意されているので、因子分析とは所詮トートロジー (ある理論で説明されている現象を、別の理論であたかもそれが新たに現象を説明したかのように述べること) でありえることも気に留めておくとよい。

```
# kodomo.txt がワークディレクトリに保存してある場合
> chr.dat <- read.table("kodomo.txt", header=TRUE)

# データセットの中身を確認する
> names(chr.dat)
```

[*1] 著者の web サイトよりダウンロードすることができる。

```
[1] "Q1"     "Q2"     "Q3"     "Q4"     "Q5"
[6] "Q6"     "Q7"     "Q8"     "desire"

# factanal() には最低でもデータが格納されているオブジェクトと
# 因子数 factors を指定する必要がある。
# 回転方法 rotation にはプロマックス法を指定した。
# scores は因子得点を計算するために指定する。
# 因子負荷量は最尤法によって計算される。
> result <- factanal(chr.dat[,1:8], factor=2,
+ rotation="promax", scores="regression")

# result に格納されているオブジェクトを確認する
> names(result)
 [1] "converged"    "loadings"     "uniquenesses" "correlation"
 [5] "criteria"     "factors"      "dof"          "method"
 [9] "scores"       "STATISTIC"    "PVAL"         "n.obs"
[10] "call"
```

5.2.3　独自性と共通性

因子分析を行うための関数 factanal() はいくつかのオブジェクトを返してくれる。このオブジェクトとは、色んな結果がそれぞれの箱に入れられている、この箱だと考えればよい。そのうちの 1 つに uniquenesses というオブジェクトが result のなかに格納されていることが確認できるだろう。

これは独自性と呼ばれるものであるが、上記の因子分析モデルに含まれる個々の回帰モデルの誤差 e_i の分散のことである（これを**誤差分散**という）。この**独自性**は、因子分析モデルにおける個々の回帰モデルにおいて、説明変数である因子 (潜在変数) が応答変数である観測変数を説明しきれていない量のことを表している。つまり、この値が 0 であるということは、説明変数である因子が応答変数である観測変数を完全に説明できていることを意味するのである。

独自性は以下のようにコマンドすることで得られる。

```
> result$uniquenesses
       Q1        Q2        Q3        Q4
```

```
0.7661545 0.7461025 0.6329522 0.7420259

       Q5        Q6        Q7        Q8
0.7291654 0.7782045 0.6373003 0.6270489
```

一方で**共通性**という値もあり、これは独自性とは逆に、因子が応答変数を説明できている量を表すものである。そして共通性は、独自性を ζ とすると、共通性 η は次式によって計算される。

$$\eta = 1 - \zeta \tag{5.8}$$

これは R で以下のようにして求めることができる。

```
> 1 - result$uniquenesses
       Q1        Q2        Q3        Q4
0.2338455 0.2538975 0.3670478 0.2579741

       Q5        Q6        Q7        Q8
0.2708346 0.2217955 0.3626997 0.3729511
```

まとめると、独自性と共通性の和は必ず 1 になるので、独自性よりも共通性の値の方が大きいほど、それは因子が応答変数を上手く説明できているということを意味しているのである。この共通性という値は、より高度に因子分析の結果を読み解くための鍵となっているのだが、ここではこれ以上の説明は省略することにする。

5.2.4 因子負荷量

因子負荷量とは回帰分析モデルでいうところの回帰係数のことであり、上記の因子分析モデルに含まれる個々の回帰モデルの β_i を推定した値である。

これは次のようにして得ることができる。

```
> result$loadings

Loadings:
```

```
   Factor1 Factor2
Q1 -0.197   0.416
Q2          0.508
Q3          0.611
Q4          0.512
Q5  0.467  -0.176
Q6  0.475
Q7  0.607
Q8  0.616

                Factor1 Factor2
SS loadings       1.247   1.115
Proportion Var    0.156   0.139
Cumulative Var    0.156   0.295
```

心理学の分野においては、慣例的に因子負荷量の絶対値が 0.4 以上である変数がその因子に含まれる変数であると判断されることが多い。したがって、Q_1, Q_2, Q_3, Q_4 は第 2 因子 (Factor2) に含まれ、Q_5, Q_6, Q_7, Q_8 は第 1 因子 (Factor1) に含まれるといえる。繰り返すが、これは元々、このようになるように質問項目を作成したのだから最初から予想できた結果である。

ところで、出力結果には SS loadings とあるが、これは**固有値**と呼ばれるものである。探索的に因子分析を行う場合、因子数がいくつであるとするのが妥当か分からないことが多い。そこで、手順としては最初に多めの因子数を指定しておき、この固有値や Cumulative Var という値を参考に因子数を決めるのである。

一応、固有値の値の目安としては 1.00 以上であることが望ましいとされているが、固有値をプロットして判断することが適切であろう。なぜならば、たとえ固有値が 1.00 以上であっても、いくつかの因子の固有値がほとんど 1.00 に近い値を持っているようならば、そのいずれも採用することになってしまうからである (それでは余計に因子数を増やすことになってしまう)。

もう 1 つの判断指標としては**累積寄与率** (Cumulative Var) がある (Proprtion Var は各因子の寄与率)。実は心理学の分野では、この指標はあまり役立たない (判断の基準として採用されていない) もので、実務的にあまり有用ではない。一般に累積寄与率は 80% 以上であることが好ましいとされているが、心理学の分野でそのような値が得られることはほとんどないからである。

5.2.5 因子得点

最後に因子得点について説明しておこう。この**因子得点**とは、文字通りある因子がもつ得点のことである。例えば、今回の解析で想定した「メリット」という因子 (潜在変数) のもつ値のことが因子得点である。元々、「メリット」という変数は存在しない変数なので、この変数の値は観測変数と因子の線形結合で算出 (推定) されることになる。

要するに、因子得点は因子分析をすることによって、新たに作られた観測変数であると考えればよい。ただし、因子分析モデルの当てはまりが悪い場合は、この因子得点の推定値も信頼できるものではないので、使用には注意が必要である。そういう観点から、心理学ではある因子に含まれる変数同士の合計得点を尺度得点として分析に用いることが多い。

R で因子得点を得るためには、必ず `factanal()` の引数に `scores = "regression"`と指定しておかなければならない。そして因子得点を抽出して出力させるためには次のようにコマンドする。

```
> f.scores <- result$scores
```

ここで `f.scores` には、1 列目に第 1 因子の因子得点が、2 列目に第 2 因子の因子得点が格納されている。つまり、第 2 列の因子得点は「メリット」という変数、第 1 列の因子得点が「デメリット」という変数として分析に扱うことができるわけである。試しに、これらを説明変数とし、`chr.dat` に含まれる `desire`(子供が欲しいという願望度) を応答変数とした重回帰モデルを解析してみよう。

```
> summary(mr.model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.5909     0.2637  210.78   <2e-16 ***
merit         4.2164     0.3384   12.46   <2e-16 ***
demerit      -4.2742     0.3289  -12.99   <2e-16 ***
```

この結果を見ると、メリットを多く感じている人は子供が欲しいという願望度が高くなり、逆にデメリットを強く感じている人は願望度が低くなるだろうということが分かる。これとは別の方法で、尺度得点を計算し、それを説明変数として解析した場合とを比べてみよう。

```
> # メリットの尺度得点
> ms.score <- apply(chr.dat[,1:4], 1, sum)
> # デメリットの尺度得点
> ds.score <- apply(chr.dat[,5:8], 1, sum)

> mr.model2 <- lm(DESIRE ~ ms.score + ds.score)
> summary(mr.model2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.21003    1.55272   35.56   <2e-16 ***
ms.score     1.03769    0.08365   12.40   <2e-16 ***
ds.score    -1.04780    0.08427  -12.43   <2e-16 ***
```

これら 2 つの方法によって解析した場合、どちらがより良く実測値を予測できているだろうか。それを確かめる方法はいくつかあるが、最も単純で分かりやすい方法は実測値と予測値の相関係数を求めてみることである。この相関係数の値が高いほど、よい予測値が得られているといえるのである。

```
> # 因子得点を用いた場合のモデル
> cor(DESIRE, predict(mr.model))
[1] 0.8113454

> # 尺度得点を用いた場合のモデル
> cor(DESIRE, predict(mr.model2))
[1] 0.8271532
```

5.3　主成分分析モデル

主成分分析は高次元のデータ (多変量データ) をより低次元の空間に表現しようとする多変量解析法の 1 つである。言い換えれば情報の集約が主な目的となる。そこでまず初めに多変量データを 1 次元上に表現するような最も単純な例を紹介する。続いて主成分分析の応用例として PCA 回帰という方法を紹介する。

5.3.1 多変量データの集約としての主成分分析

主成分分析モデルとは、潜在変数を応答変数として 2 変量以上の説明変数でそれを説明しようとする回帰分析モデルである。換言すれば、主成分分析は 2 つ以上の変数から、新たな 1 つの合成変数を作り出す手法といえる。すなわち、多変量データを集約し、より少ない次元で表現しようとするものである。

観測変数 $X_1, X_2, ..., X_i$ があったとき、新たに作られる合成変数 $Z_1, Z_2, ..., Z_i$ は次式によって表現される。

$$
\begin{aligned}
Z_1 &= w_1X_1 + w_2X_2 + ... + w_iX_i \\
Z_2 &= w\prime_1X_1 + w\prime_2X_2 + ... + w\prime_iX_i \\
&\cdot \\
&\cdot \\
&\cdot \\
Z_i &= w\prime\prime_1X_1 + w\prime\prime_2X_2 + ... + w\prime\prime_iX_i
\end{aligned}
$$

主成分分析では新たに作られる合成変数のことを特に**主成分**という（これが主成分分析と呼ばれる所以である）。また、上式が示すように、主成分分析では i 個の観測変数と同じ数の合成変数（主成分）が作成されることになる。そして、観測変数の係数 w は、回帰分析における偏回帰係数と同じものであると考えればよく、合成変数を観測変数が説明する度合い（この度合いのことを「重み」という）のことである。

ここで極端なデータセットを用いて、主成分分析を実際に行ってみることにしよう。表 5.3 は 16 人の被験者に対して、異性のパートナーを選ぶ際に重要だと考えられる 4 つの項目について、各項目 10 点満点で評価してもらったデータである。

R には主成分分析を行うために `prcomp()` と `princomp()` という 2 つの関数が用意されている。どちらも大きな違いはないが、`prcomp()` はサンプルサイズが変数の数よりも少ない場合にも適用できる、という利点をもっている。また、`prcomp()` は計算時に不偏分散を用いるのに対して、 `princomp()` では標本分散を用いている。両者に若干の違いはあるが、基本的にはどちらを使っても構わないだろう。

```
# 以下のようなデータフレームを用意しておく
> mydata
   face style talk response
1     8    10    2        1
2     8     7    1        3
3     9    10    1        2
```

表 5.3　異性パートナーに関するデータセット

被験者 No.	顔	スタイル	話し方	態度
1	8	10	2	1
2	8	7	1	3
3	9	10	1	2
4	10	10	3	2
5	9	9	3	3
6	9	10	2	1
7	8	7	1	2
8	8	9	2	1
9	1	2	10	7
10	1	2	8	8
11	2	2	9	9
12	3	1	10	9
13	2	3	8	9
14	3	3	9	10
15	4	1	7	8
16	3	2	8	8

```
4    10    10    3        2
(途中省略)

# princomp() を使って主成分分析を行う
# 引数 cor=TRUE は原データを標準化して解析することを指定するものである
> result <- princomp(mydata, cor=TRUE)

# summary() の引数に loadings=TRUE とすることで、主成分負荷量が出力される
> summary(result, loadings=TRUE)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4
Standard deviation     1.9471673 0.30051060 0.28652740 0.19009186
Proportion of Variance 0.9478651 0.02257665 0.02052449 0.00903373
Cumulative Proportion  0.9478651 0.97044178 0.99096627 1.00000000
```

```
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4
face     -0.500          0.744 -0.434
style    -0.500  0.676 -0.111  0.530
talk      0.499  0.730        -0.466
response  0.501         0.659  0.560
```

前述したように、主成分分析では観測変数の数と同じ数の合成変数が作成されることになるので、適当な数だけ合成変数を採用して、それ以降の合成変数は切り捨てる必要がある。その判断基準となるのが Standard deviation（標準偏差）と Cumulative Proportion（累積寄与率）である。

ここでいう標準偏差とは、合成変数の標準偏差のことであり、例えば、第 1 合成変数の標準偏差は 1.95 である。主成分分析では、作成される合成変数の分散が最大になるように重み（w）が計算されるので、この値が大きいほど「よい合成変数である」ということがいえる。判断の基準としては、大まかにいって 1 以上の合成変数を採用し、それ以下の合成変数を切り捨てるというのが一般的である。

一方で累積寄与率もいくつの合成変数を採用するかの判断基準として役立つ。簡単にいえば、この累積寄与率とは第 i 番目までの合成変数がデータ全体を説明している割合のことである。慣例的に 80%（0.80）という寄与率があれば、十分にデータ全体を説明していると判断される。したがって、今回の場合は第 1 番目の合成変数（Comp.1）だけでも、すでにデータ全体の 95%（0.95）を説明していることになるので、第 2 番目以降の合成変数は切り捨てればよいことになる。

さて、Loadings に記されている数値は合成変数を説明する各変数の重み（w）である。この重み（w）のことを、特に主成分負荷量というが、前述したように回帰分析における偏回帰係数に当たるものである。

これにしたがって、第 1 番目の合成変数（第 1 主成分）は次のように表される。

$$Comp.1 = -0.500face - 0.500style + 0.499talk + 0.501attitude \tag{5.9}$$

新たに作られた合成変数 $Comp.1$ はどういった変数なのか分析者には分からない（そういう意味でこれを潜在変数と考えることができる）。主成分分析を行う際には、しばしば分析者がこの合成変数にラベルをつけることがある。例えば、「異性の判断基準値」などというラベルをつけることができるかもしれないし、「離婚の可能性」というラベルをつけることができるかもしれない。

こういった合成変数のラベル付け（ラベリング）は分析者の主観によるものなので、どれほど妥当性のあるものであるか、なかなか難しいところである。というのも、原データが少

し異なっているだけでも、全く異なった結果が得られることもあるので、あるケースでは全く別のラベリングがなされる可能性もあるのからである。いいかえれば、このケースでのラベリングは別のケースでは全く意味を持たないこともありえるということである。

それでは第 1 主成分の解釈はどのようになるかを考えてみよう。変数 $face$（顔）と $style$（スタイル）はいずれもマイナスの符号を持っているので、$Comp.1$ に対して負の効果をもたらしていることが分かる。逆に変数 $talk$（会話）と $attitude$（態度）はプラスの符号を持っているので、$Comp.1$ に対して正の効果をもっているといえる。そしてこれら 4 つの変数の効果量はどれも同じ程度であることが分かる（どの変数の係数も約 0.50 である）。なお、4 つの変数の効果量を比較できるのは、標準化解だからである。

仮に $Comp.1$ を「異性の判断基準値」としたならば、$face$ や $style$ のように外見を重視する人は $Comp.1$ の値はマイナス方向に大きな値をとるといえるだろう。それに対して $talk$ や $response$ といった内面的な要素を重視する人はプラス方向に大きな値をとるだろうということが分かる。

そこで主成分得点を求めてプロットしてみると、その傾向が顕著に見てとれる（図 5.3）。主成分得点とは、回帰分析でいうところの予測値のようなものである。例えば、被験者 No.1 のデータ $\{face, style, talk, attitude\} = \{8, 10, 2, 1\}$ の平均偏差値を第 1 合成変数の式に代入することによって被験者 No.1 の主成分得点が得られる。

ただし、これは原データを標準化せずに主成分分析を行った場合にいえることである。今回の例では原データを標準化して行っているので、代入する値も標準化されたデータの平均偏差値でなくてはならない。

R では次のようにして簡単に主成分得点を得ることができる（今回、見るべきなのは `Comp. 1` の主成分得点のみである）。

```
> result$scores
         Comp.1       Comp.2       Comp.3      Comp.4
 [1,] -2.083127  0.178410538 -0.345297966  0.06772526
 [2,] -1.520295 -0.564240732  0.140169589  0.09191751
 [3,] -2.233419  0.009637521  0.085185142  0.23119187
 [4,] -2.101801  0.456844578  0.298651584 -0.16950473
（途中省略）
```

`Comp. 1` の主成分得点の様子を `barplot()` で表してみると、各被験者の「異性の判断基準値」がよく分かる。どうやら被験者 1 から 8 は外見を重視するタイプで、被験者 9 から 16 は内面的な要素を重視するタイプだということがいえそうである。

```
> barplot(result$scores[,1])
```

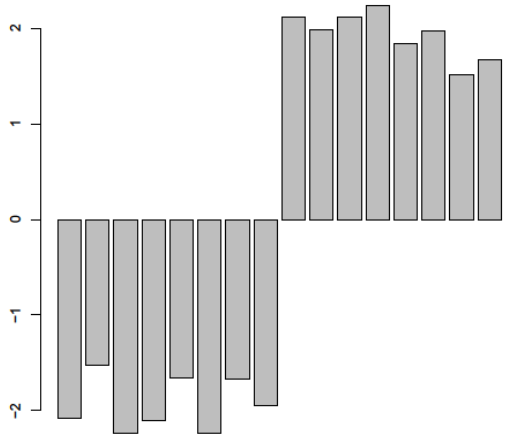


図 5.3 主成分得点のプロット

5.3.2 PCA 回帰

主成分分析のことを英語で Principal Component Analysis というが、この頭文字をとって PCA と称されることがある。既に紹介したこの主成分分析 (PCA) によって得られた合成変数 (主成分) を説明変数として用い、ある応答変数を説明しようとする分析法を PCA **回帰**という。

ここで表 5.3 のデータセットに、新たに結婚願望度という変数を加えたものを表 5.4 として用意する。このようなデータが得られている場合、結婚願望 (*desire*) を応答変数として、他の変数を説明変数とした重回帰モデルを解析する方法が考えられる。しかし、各変数同士の相関係数を確認してみると、説明変数同士でかなり高い相関があることが見てとれる。

```
> # 新たなデータセットが isei.txt としてワークディレクトリに
> # 保存されているとする
> isei <- read.table("isei.txt", header=TRUE)
> names(isei)
[1] "face"     "style"    "talk"     "attitude" "desire"
```

```
> cor(isei)   # 相関行列の出力
              face      style       talk   attitude     desire
face       1.0000000  0.9378874 -0.9337159 -0.9181890  0.7711699
style      0.9378874  1.0000000 -0.9098850 -0.9431195  0.8828999
talk      -0.9337159 -0.9098850  1.0000000  0.9401025 -0.7809968
attitude  -0.9181890 -0.9431195  0.9401025  1.0000000 -0.8529417
desire     0.7711699  0.8828999 -0.7809968 -0.8529417  1.0000000
```

このような場合において、重回帰分析を行うと**多重共線性**が生じてしまうことがある。主成分分析は各合成変量の分散が最大になるように (つまり各変数同士の相関が 0 になるように) 作られるので、主成分得点を用いることは、元の変数同士に強い相関が見られるような場合に有効である。

表 5.4　異性パートナーに関するデータセット 2

被験者 No.	顔	スタイル	話し方	態度	結婚願望
1	8	10	2	1	34
2	8	7	1	3	30
3	9	10	1	2	36
4	10	10	3	2	41
5	9	9	3	3	37
6	9	10	2	1	38
7	8	7	1	2	33
8	8	9	2	1	35
9	1	2	10	7	23
10	1	2	8	8	30
11	2	2	9	9	21
12	3	1	10	9	19
13	2	3	8	9	25
14	3	3	9	10	15
15	4	1	7	8	12
16	3	2	8	8	12

それでは実際に重回帰モデルと PCA 回帰モデルの両方を解析してみることにしよう。以下にその解析例を示す。

```
> pca.result <- princomp(isei[,1:4], cor=TRUE)
> names(pca.result)
[1] "sdev"     "loadings" "center"   "scale"
[5] "n.obs"    "scores"   "call"
> pca.score <- pca.result$scores[,1]

> # 主成分得点を用いて PCA 回帰を行う
> pca.reg <- lm(attitude ~ pca.score, data=isei)
> summary(pca.reg)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1875     0.1979   26.21 2.68e-13 ***
pca.score     1.7113     0.1017   16.84 1.10e-10 ***

> # 通常通り重回帰分析を行う
> nr.reg <- lm(desire ~ face + style + talk + attitude, data=isei)
> summary(nr.reg)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.6398    12.6745   2.023   0.0681 .
face         -1.5402     1.2742  -1.209   0.2521
style         2.7360     1.1777   2.323   0.0403 *
talk          0.1797     1.1763   0.153   0.8813
attitude     -1.0791     1.3022  -0.829   0.4249
```

この結果からいえることは、まず通常通りに重回帰分析を行った場合は *style* 以外の変数は全て予測に使えない変数ということになる。ところが PCA 回帰ではこれら 4 つ全ての変数の情報をほぼ余すことなく使用できている点に PCA を使う利点がある。また、説明変数の数が重回帰分析のときよりも少なくなるので、解釈をするうえでも結果をより単純なものとして考えることができる。

もっとも、この場合はもっと根本的な問題として、多重共線性が生じているので、そもそも重回帰分析を行うことが適切ではないのである。相関行列を見ると、応答変数である *desire* と説明変数である 4 つの変数いずれとも高い相関がある。つまり、*desire* を説明するために

どの変数も重要である。ところが、重回帰分析の結果では重要な変数は $style$ だけということになっている。これはある変数 (例えば $face$ という変数) が別のある変数 (例えば $talk$ という変数) の説明量を減少させているためである (この多重共線性については、より高度な議論が可能であるが、本書ではこの辺りで留めておくことにする)。

5.4　判別分析モデル

判別分析とは既存のデータを基に未知のサンプル（被験体）がどの群に属するかを予測するための分析手法である。未知のサンプルがどの群に属するかを予測するための判別ルールを作るわけであるが、ここでは**線形判別分析**といって直線によって判別する方法を主に紹介する。さらに、判別分析と関連のある多変量分散分析、重回帰分析、ロジスティック回帰分析 (2 値ロジットモデル) を併せて紹介する。

5.4.1　2 群の判別分析

線形判別関数による判別

まずはもっとも単純な 2 **群の判別分析**の例を示す。表 5.5 のデータは男女の身長と体重のデータであり、仮に未知のサンプル（被験体）の身長と体重のデータだけが得られていたとする。このとき、No. 11 のサンプル（被験体）は性別が不明であるので、判別分析によって No. 11 のサンプルの性別を予測することを考える。

表 5.5　身長・体重と性別のデータ

No.	身長（Height）	体重（Weight）	性別（Sex）
1	177	75	male
2	180	73	male
3	175	70	male
4	182	85	male
5	170	69	male
6	166	65	female
7	159	60	female
8	160	61	female
9	155	58	female
10	163	68	female
11	170	67	?

判別分析のモデル式は次のように表すことができる。

$$Sex = \alpha + \beta Height + \beta Weight \tag{5.10}$$

ここで従属変数 Sex は 2 つのカテゴリをもつカテゴリカル型の変数であり、独立変数 $Height$ と $Weight$ は連続型の変数である。勘のいい人は気づいたかもしれないが、これは 2 値ロジットモデルとして解析できるモデルである。実際、表 5.5 のデータセットに対してロジスティック回帰分析を行っても問題なく、得られる結論も判別分析と似た部分がある。これについては後述することにして、とりあえず判別分析モデルとして解析してみよう。

R で判別分析を行うには MASS パッケージに含まれている `lda()` という関数を用いる。

```
# MASS パッケージの呼び出し
> library(MASS)

# 身長・体重・性別のデータを用意する
> Height <- c(177, 180, 175, 182, 170, 166, 159, 160, 155, 163)
> Weight <- c(75, 73, 70, 85, 69, 65, 60, 61, 58, 68)
> Sex <- rep(c("male", "female"), c(5, 5))
> Sex <- factor(Sex, levels=c("male", "female"))

# lda() を用いて判別分析を行う
> disc.model <- lda(Sex ~ Height + Weight)

# 結果の出力
> disc.model
Call:
lda(Sex ~ Height + Weight)

Prior probabilities of groups: # 事前確率
  male female
   0.5    0.5

Group means: # 群平均
       Height Weight
male    176.8   74.4
female  160.6   62.4
```

```
Coefficients of linear discriminants: # 線形係数
              LD1
Height -0.3114300
Weight  0.0986707
```

さて、Prior probabilities of groups:という部分は事前確率、つまり $male$ と $female$ に属する確率は等しく 0.5 であると仮定した、ということである。これは引数 prior で指定できるが、省略すると R 側が勝手に計算してくれる。今回の場合だと、事前に分かっている $male$ の割合は 5 / 10 = 0.5 であり、$female$ の割合は 5 / 10 = 0.5 であるから、その値が Prior probabilities of group:として出力されているわけである。

一方、Group means:は群平均であり、$male$ について身長の平均値は 176.8、体重は 74.4 である。$female$ について身長の平均値は 160.6、体重は 62.4 であることを示している。

さらに Coefficients of liner discriminants:は線形係数、要するに回帰分析でいうところの偏回帰係数のことである。判別分析のモデル式 (式 5.10) の β_1 と β_2 を推定した値のことである。ここでは $Height$ の係数は $\beta_1 = -0.311$ であり、$Weight$ の係数は $\beta_2 = 0.099$ であることが分かる。

定数項 α は次のようにして計算することができる（少々面倒だが、自分で計算しなければならない)。

```
# 群平均のデータ行列を group.mean に代入
> (group.mean <- disc.model$mean)
       Height Weight
male    176.8   74.4
female  160.6   62.4

# 線形係数のデータ行列を coef に代入
> (coef <- disc.model$scaling)
              LD1
Height -0.3114300
Weight  0.0986707

# 行列のかけ算は %*% で行える
# その結果を a に代入
> (a <- group.mean %*% coef)
```

```
             LD1
male   -47.71973
female -43.85861

# 行列 a の列平均を計算する
# 引数に 2 ではなく 1 を指定すれば行に対して計算される
# 仮に mean を sum にすると、平均ではなく合計が計算される
> apply(a, 2, mean)
      LD1
-45.78917

# 以上の手順は 1 回で済ますこともできる
> apply(disc.model$means %*% disc.model$scaling, 2, mean)
      LD1
-45.78917
```

以上の作業から、次式のような線形判別式が得られる。

$$Sex = -(-45.789) - 0.311Height + 0.099Weight \tag{5.11}$$

この判別式の $Height$ と $Weight$ それぞれにあるサンプル（被験体）のデータを代入することによって判別得点が得られる。そして、その判別得点がプラスの値かマイナスの値かによってそのサンプル（被験体）がどちらのグループに属するかを判別するというわけである。

```
# 線形判別関数を定義する
> f <- function(H, W) -(- 45.789) - 0.311*H + 0.099*W

# 引数に Height と Weight を指定すれば、各サンプルの判別得点が計算される
> f(Height, Weight)
 [1] -1.833 -2.964 -1.706 -2.398 -0.250
 [6] 0.598  2.280  2.068  3.326  1.828
```

これは predict() を利用して簡単に求めることができる（丸め誤差の関係で上記の値とは若干異なる）。

```
# 判別得点の出力
> predict(disc.model)$x
          LD1
1  -1.9336427
2  -3.0652743
3  -1.8041362
4  -2.5040859
5  -0.3456566
6   0.5053807
7   2.1920375
8   1.9792782
9   3.2404163
10  1.7356830
```

表 5.5 のデータセットをプロットしたものに先ほど求めた判別式が描く直線を加えると図 5.4 のようになる（データポイントの m は男性を、f は女性を表している）。これを見ればよく分かるが、あるサンプル（被験体）のデータポイントが直線よりも上側にあれば（あるサンプル（被験体）の判別得点がプラスであれば）$male$ であると判別される。逆に下側（判別得点がマイナス）であれば $female$ であると判別される。

ここで No. 11 のサンプルのデータ（$Height = 170, Weight = 67$）を判別式に代入して、このサンプルがどちらのグループに属するかを予測してみる。また、No. 11 のサンプルを図示するには points() を使えばよい (図 5.5)。

```
# 先ほど定義した判別関数の中身を確認
> f
function(H, W) -(- 45.789) - 0.311*H + 0.099*W

# 身長 170、体重 67 として判別得点を求める
> f(170, 67)
[1] -0.448

# No.11 のサンプルを新たに書き加える
> points(67, 170, pch=17, cex=1.5)
```

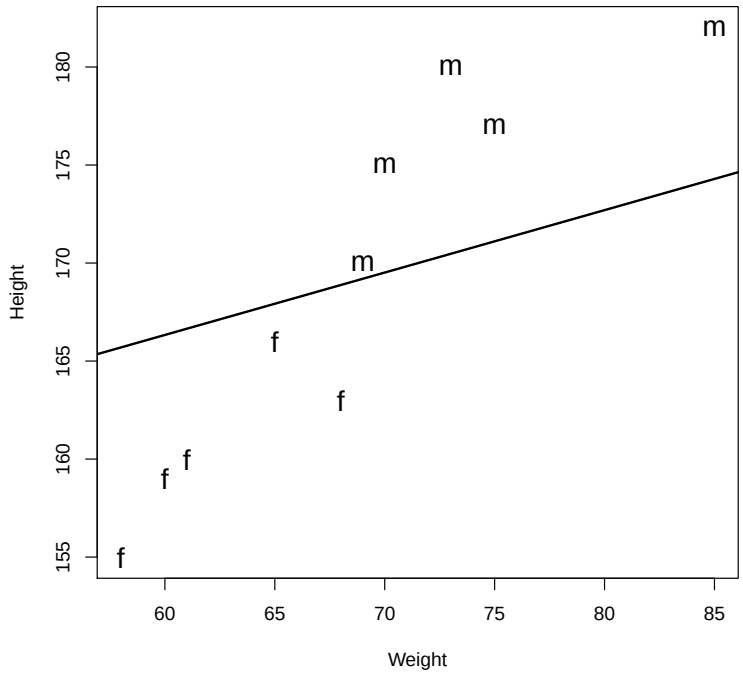


図 5.4 線形判別関数による 2 群の判別

結局、未知のサンプルは $male$ であることが予測されたということが見てとれる。しかし、このように図示できるのは独立変数の数が 2 つの場合のみである。独立変数が 3 つ以上あるときは頭の中で想像するしかないのである。もっとも 3 つならば 3 次元プロットができるが、見やすさの観点からいってもあまりおススメはできるものではない。

ところで、図 5.4 で描かれた直線の方程式はどのようにして導かれたものであるかを説明しておこう。そもそもパラメータ β_1 および β_2 は次の連立方程式を解くことによって得られるものである (式 5.12)。

$$\begin{cases} \alpha + \beta_1 Height + \beta_2 Weight = 0 \\ Sex = \alpha + \beta_1 Height + \beta_2 Weight \end{cases} \tag{5.12}$$

前者は 2 つのグループを分割する直線の方程式で、後者は線形判別関数である。そして定数項である α は、群平均の中点を通るように求められることになる。

この 2 つのグループを分割する直線の方程式について、α と $\beta_2 Weight$ を右辺に移項すると、式 5.13 のように変形されることになる。

$$\beta_1 Height = -\alpha - \beta_2 Weight \tag{5.13}$$

さらに $Height$ の係数を右辺に移項すると、式 5.14 のようになる。この式 5.14 が図 5.4 に描かれている直線の方程式である。

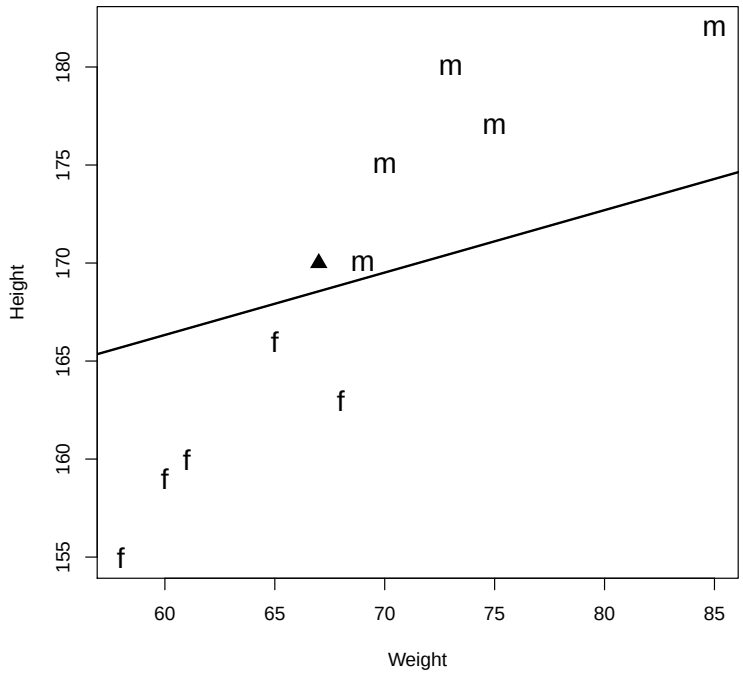


図 5.5　No.11 のサンプルを書き加えたグラフ

$$Height = -\frac{\alpha}{\beta_1} - \frac{\beta_2}{\beta_1} Weight \tag{5.14}$$

線形判別分析と重回帰分析

2 群の判別分析と重回帰分析には関連がある。試しに `lm()` を使って解析して、そこから得られた直線の方程式がどのような直線であるかを図示してみよう (図 5.6)。

```
# グラフを新しく描き直す
> plot(Weight, Height, och=c("m", "f")[Sex], cex=1.5)
> points(x, f2(x), type="l")

# 性別を表す変数を連続型として用意する
> Sex2 <- rep(c(1, 0), c(5, 5))
> Sex2
 [1] 1 1 1 1 1 0 0 0 0 0

# lm() を使って分析する
```

```
> lm.model <- lm(Sex2 ~ Height + Weight)
> lm.model

Call:
lm(formula = Sex2 ~ Height + Weight)

Coefficients:
(Intercept)       Height       Weight
   -9.26337      0.06640     -0.02104

# 式の変形は上でやったとおり
> lm.f <- function(x) (-9.263 / -0.066) + (-0.021 / -0.066)*x
> points(x, lm.f(x), type="l", lty=2, lwd=2)

> legend(60, 180, # 凡例を書く座標
+         c("mra.line", "lda.line"), # 凡例のラベル
+         lty=c(1, 2), # 凡例のラインの色
+         lwd=2 # ラインの太さ
+ )
```

図 5.6 を見れば明らかだが、2 つの直線は切片（つまり定数項）が異なるだけで、傾きは同じである。つまり、2 つの平行な直線が描かれるかたちになっている。これはどういうことかというと、重回帰によって得られた直線を、male と female の群平均の中点を通るように切片を調整しているということである。

具体的に例を示そう。

```
> disc.model$mean
       Height Weight
male    176.8   74.4
female  160.6   62.4
```

これは $male$ と $female$ の群平均である。$male$ については $(Weight, Height) = (74.4, 176.8)$ という座標に示される。$female$ については $(Weight, Height) = (62.4, 160.6)$ という座標に示される。この 2 つの座標を図示し、さらにその 2 点を直線でつないでみると

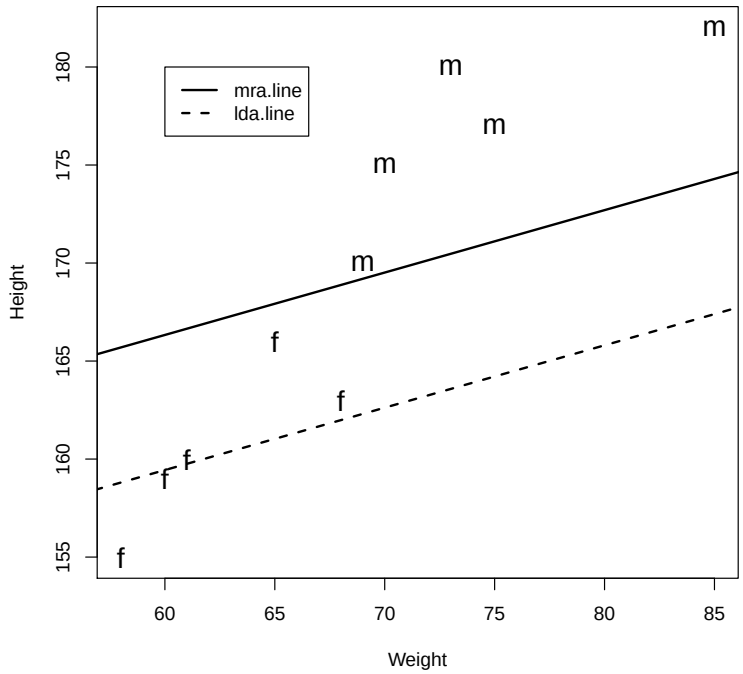


図 5.6　判別分析の直線と重回帰分析の直線

図 5.7 のようになる。

```
# 男性の座標
points(74.4, 176.8, pch=19, cex=1.5)

# 女性の座標
points(62.4, 160.6, pch=19, cex=1.5)

# 座標 (74.4, 176.8) と (62.4, 160.6) を結ぶ直線を描く
segments(74.4, 176.8, 62.4, 160.6, lty=3, lwd=2)
```

このように、2 群の判別分析とは重回帰分析によって得られた直線の方程式を各群の群平均の中点を通るように定数項を調節したものである。つまり、重回帰分析のプログラムがあれば 2 群の判別分析が可能であるということである。

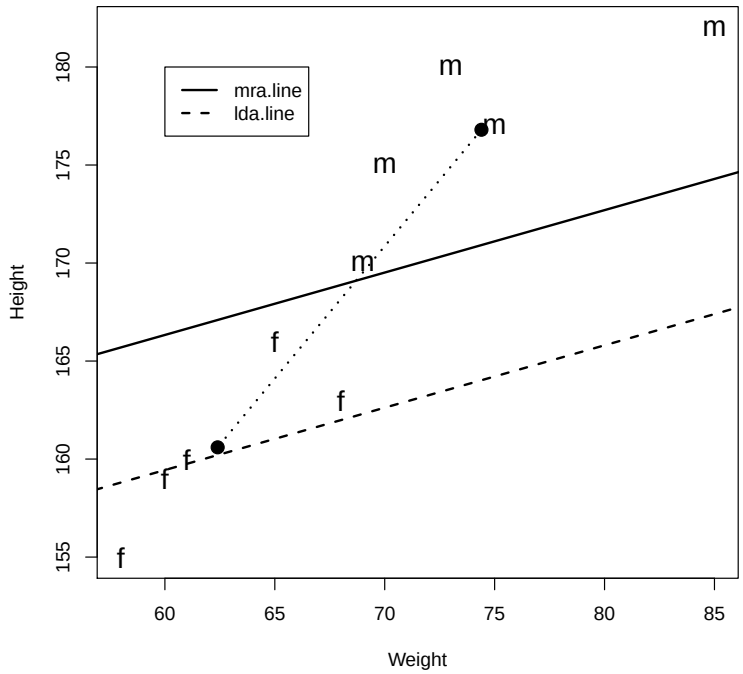


図 5.7 男性の群平均と女性の群平均を結ぶ直線

5.4.2 判別分析と多変量分散分析

実は判別分析と多変量分散分析は似たような性質をもつ分析法である。表 5.5 について、判別分析は Sex を応答変数とし、$Height$ と $Weight$ を説明変数としていた。それに対して多変量分散分析は $Height$ と $Weight$ を応答変数、Sex を説明変数としたモデルである。つまり、全く同じ形式のデータセットに対して、この 2 つ異なるアプローチが可能なのである。

判別分析モデルのモデル式 $Sex = Height + Weight$ は「$Height$ と $Weight$ によって、Sex を判別できるか」という問題を取り扱っている。一方で多変量分散分析モデルのモデル式 $Height, Weight = Sex$ は「Sex は $Height$ と $Weight$ を説明できるか」という問題を取り扱うことになる。これは換言すれば、2 つの応答変数に男性と女性の平均値をそれぞれ当てはめたとき、それは 2 つの応答変数の変動を上手く説明できるか、ということである。つまり、説明の方法が逆方向であるだけで、結果としては同じことをしているにすぎないのである。

試しに関数 `manova()` を用いて、表 5.5 のデータセットを解析してみよう。

```
> Sex
 [1] male   male   male   male   male
```

```
 [6] female female female female female
Levels: male female
> Height
 [1] 177 180 175 182 170 166 159 160 155 163
> Weight
 [1] 75 73 70 85 69 65 60 61 58 68

> manova.result <- manova(cbind(Height, Weight) ~ Sex)
> summary(manova.result, test="Wilks")
          Df   Wilks approx F num Df den Df  Pr(>F)
Sex        1 0.17672   16.306      2      7 0.00232 **
Residuals  8

> summary(aov(manova.result))
 Response Height :
            Df Sum Sq Mean Sq F value    Pr(>F)
Sex          1  656.1   656.1  33.646 0.0004049 ***
Residuals    8  156.0    19.5
---

 Response Weight :
            Df Sum Sq Mean Sq F value   Pr(>F)
Sex          1  360.0  360.00  12.610 0.007499 **
Residuals    8  228.4   28.55
```

この結果が示す事実は、$Height$ も $Weight$ も Sex によって有意に説明されるということである。もう少しいえば、$Height$ に男性と女性の平均値を当てはめたとき、$Height$ の変動を上手く説明できるということである。言い方を変えれば、$Height$ のデータのばらつきは男性と女性という 2 群に分離できるということである（$Weight$ についても同じである)。

5.4.3　判別分析モデルとロジスティック回帰分析

先に述べたように表 5.5 のデータセットは応答変数が 2 値データであるので、これはロジスティック回帰分析 (2 値ロジットモデル) として扱うことも可能である。そこで、ここでは実際に 2 値ロジットモデルとして解析した場合はどのような結果が得られるのかを試してみよう。

2 値ロジットモデルを扱うには R の glm() を用いるが、表 5.5 のデータセットは使うことができない。なぜかというと、glm() では、ある説明変数が完全に応答変数を説明してしまうようなデータセットを受けつけないからである。試しに実行してみると以下のようなエラーメッセージが出てしまう。

```
> logit.model <- glm(Sex ~ Height + Weight, binomial)
 警告メッセージ：
In glm.fit(x = X, y = Y, weights = weights,
start = start, etastart = etastart,  :
   数値的に 0 か 1 である確率が生じました
>
```

そこで今回は練習のために、少しデータセットを改変して性別を表す新たな変数 $new.Sex$ を用意することにしよう。

```
> new.Sex <- c(1, 1, 0, 1, 1, 1, 0, 0, 0, 0)
> new.Sex <- factor(new.Sex, levels=c(1, 0),
+ labels=c("male", "female"))
> new.Sex
 [1] male   male   female male   male   male
 [6] female female female female
Levels: male female
```

これについて、glm() を用いたロジットモデルと lda() を用いた線形判別分析モデルとの結果を比べてみよう。

```
# ロジットモデルとして解析
logit.model <- glm(new.Sex ~ Height + Weight, binomial)

# 線形判別分析モデルとして解析
d.model <- lda(new.Sex ~ Height + Weight)
```

ロジットモデルについて、関数 fitted() を用いることによって各サンプルが「男性であ

るか否かという確率」が得られる (predict() の引数に type="response"としてもよい)。判別分析モデルについても、predict() をによって各サンプルが「男性のグループに属する確率と女性のグループに属する確率」が得られる。

```
> fitted(logit.model)
         1          2          3          4          5
0.10345534 0.09279978 0.23665446 0.01376522 0.41361913

         6          7          8          9         10
0.68295867 0.91941496 0.89616468 0.96317745 0.67799031

> predict(d.model)$posterior
         male     female
1  0.88262112 0.11737888
2  0.93518131 0.06481869
3  0.81321691 0.18678309
4  0.96575521 0.03424479
5  0.57647561 0.42352439
6  0.33660726 0.66339274
7  0.08549437 0.91450563
8  0.10685845 0.89314155
9  0.03484643 0.96515357
10 0.21208160 0.78791840
```

ロジットモデルの解析の結果より、例えば、サンプル 1 が男性でない確率は 0.10（10%）であることがわかる。それに対して判別分析モデルの解析結果からは、サンプル 1 が男性でない確率（女性のグループに属する確率）は 0.11（11%）と両者の結果はそれなりに一致している（少なくとも今回のケースにおいては)。

要するに（大雑把にいえば）以下の 2 つの出力（lg.prob と d.prob）は同じことを意味しており、両者の相関係数を計算すると 0.994 と非常に高い値が得られる。これはロジットモデルによって得られる予測確率と判別分析によって得られる予測確率がかなり一致していることを意味している。

```
# 女性である確率　ロジットモデル
```

```
> lg.prob <- fitted(logit.model)

# 女性である確率　判別分析
> d.prob <- predict(d.model)$posterior[,2]

# 両者の相関係数
> cor(lg.prob, d.prob)
[1] 0.993928
```

第 6 章

統計学の基礎知識

6.1 分散

6.1.1 平均値

標本とはデータの集まりと考えればよいが、標本データを代表するような値のことを代表値という。これは読者諸君も知っている平均値のことだと考えればよい。平均値といってもいくつか種類があるが、単に平均値といった場合には算術平均、あるいは相加平均と呼ばれているもののことを指している。この平均値という代表値は次式によって計算することができる。

$$mean = \frac{\sum_{i=1}^{n} x_i}{n} \tag{6.1}$$

さて、$mean$ とは「平均値」という意味である（統計学では平均値のことを average とはいわず、mean というのが一般的である）。n はサンプルサイズ（データ数）で、x_i は i 番目のデータのことを表しています。例えば、ある標本 A は {10, 12, 6, 9, 10} というデータの集まりであるとします。この標本 A の平均値は以下のようにして求められる。

$$mean = \frac{10 + 12 + 6 + 9 + 10}{5} = 9.4$$

なお、今回得られている標本は 1 つなので「標本数は 1」で、この標本を構成しているデータ数は 5 つなので「標本の大きさは 5」である。標本の大きさは**サンプルサイズ**ともいうが、よく標本数と標本の大きさ（サンプルサイズ）を混同している者がいるで気をつけるべきである。これは統計学の専門家と称されている者でさえも誤って使っている場合があるので、参考書を読むときなどには特に注意しなければならない。

6.1.2　分散

統計学において代表値は解釈上でも重要な意味を持つが、それ以上にデータのバラつきを表す指標、すなわち散布度について考えることが重要となる。この散布度というのは、要するに各々のデータが平均値からどれほど離れているか、ということを考えることに違いない。

ここで横軸に標本データの番号を、縦軸に標本データの値をとってプロットしてみる。データを図示することをプロットといい、このような図のことを散布図という。さらにこの散布図に先ほど求めた平均値を表す線を描いてみる（図 6.1）。

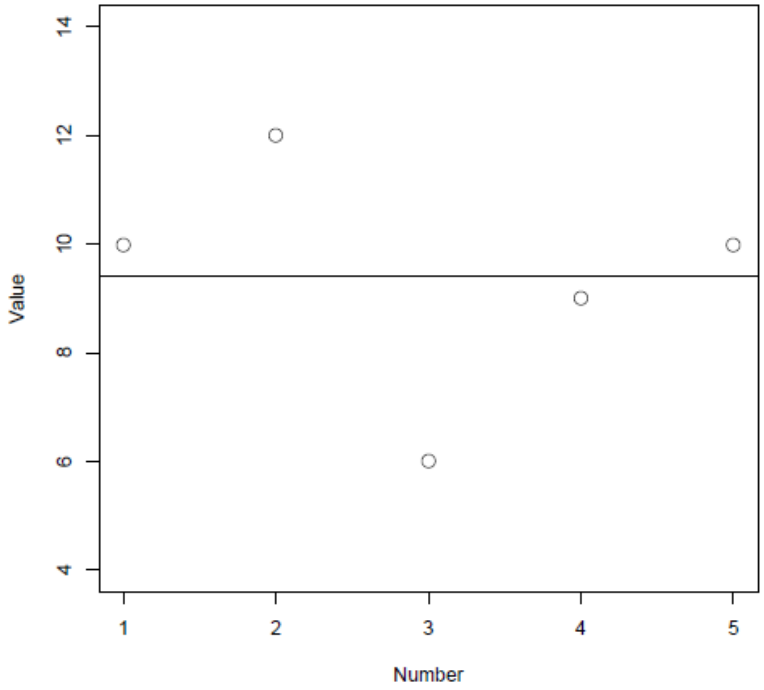


図 6.1　原データのプロットと平均値の直線

問題は平均値からの離れ具合をどのような指標で表すかということである。まず考えつくであろうことは「平均値からの差をとり、それらを合計したもの」を使うことだろう。

標本データは $\{10, 12, 6, 9, 10\}$ であり、平均は 9.4 であるから、平均からの差は：

$$\{(9.4-10), (9.4-12), (9.4-6), (9.4-9), (9.4-10)\}$$

として計算され、これらを合計すると：

$$-0.6 + (-2.6) + 3.4 + 0.4 + (-0.6) = 0$$

となり、結局は 0 になってしまう。これはどのようなデータ群であっても 0 になることが数学的に証明することができる（ここでは省略するが、文献 [16] を参照すればよいだろう）。

この平均からの差のことを、統計の専門用語では残差、あるいは偏差といい、これらの合計を残差和という。

さて、どんな場合でも 0 になってしまうのでは残差和はバラつきを表す指標として使えない。そこで次に考えるとすれば残差をそのまま合計するのではなく、残差を自乗して合計するという方法が考えられるだろう。

$$-0.6^2 + (-2.6)^2 + 3.4^2 + 0.4^2 + (-0.6)^2 = 19.2$$

これはなかなか以ってうまい具合にバラつきを表す指標として使えそうである。各残差を自乗（平方）して合計したので、これは平方和（SS, Sum of Squares）と呼ばれている。ところが、この平方和は新たなデータを加えると、それがどのようなデータであっても平方和を増加させるという性質をもっている。つまり、データ数が増えれば増えるほど、平方和の値もそれに伴って大きくなるということである。

これについての対処法は「データ数で割る」ということである。実はこの平方和をデータ数で割ったものが分散（variance）と呼ばれる、統計学で最も重要となる指標の 1 つである。分散を求める式は以下のようになる。

$$var = \frac{\sum_{i=1}^{n}(mean - x_i)^2}{n} \tag{6.2}$$

6.1.3 不偏推定量

実は式 6.2 によって計算される分散は標本分散と呼ばれるものである。これはサンプルサイズが大きい場合には母分散と近似的に一致することが知られている。しかしながら、サンプルサイズが小さくなるほど、母分散との差が大きくなってしまう。そこで、式 6.2 の分母を n ではなく、$n-1$ で割ってやることで分散の不偏推定量が得られる（これを**不偏分散**という）。

$$var = \frac{\sum_{i=1}^{n}(mean - x_i)^2}{n-1} \tag{6.3}$$

より正確にいえば、分母を $n-1$ で割るという理屈は自由度と深い関係がある。これは次章の分散分析の基本原理のところで述べることにする。ここでは、サンプルサイズが小さい場合には不偏分散を用いなければならない、ということだけ気に留めておけばよいだろう。

6.2 分散分析の原理

ここでは分散分析について詳しく説明していく。もしかしたら、ある人は「平均値を比較する分析なのに、なぜ分散分析という名前がついているのだろうか。」という疑問をもっているかもしれない。おそらく、その疑問はこのページを読み終える頃には解決しているだろう。先立って、まずは表 6.1 のような 2 群のデータセットについて考えてみよう。

表 6.1　2 群のデータセット

Group A	Group B
20	45
25	33
28	40
22	39
24	41
26	31
30	37

いきなりだが、またある人は「分散分析は 3 群以上の平均値を比べるときに使うもので、2 群の平均値の比較なら t 検定なのではないか。」と思ったかもしれない。確かにそれはその通りなのだが、そもそも分散分析とは説明変数がカテゴリカル型である場合のモデルであり、そのモデル式は次のように書ける。

$$Y = A \quad (A \text{ はカテゴリカル型})$$

このモデル式において A はカテゴリカル型の変数である。カテゴリカル型ということはいくつかの水準をもっているわけで、もし 2 つの水準をもっているならば（たとえば男性と女性といったように)、結果としてそれは 2 群の平均値について検討していることになる。もしも 3 つの水準をもってるならば（たとえば小学校、中学校、高校といったように)、それは 3 群の平均値を比較するということである。だから、一般線形モデルの立場でいうならば、古典的に扱われている t 検定も分散分析も同じ分散分析モデルという 1 つのモデルで表現されるのである。つまり、違いは単に水準数の違いだけなのである。

6.2.1　モデル式による表現

これを今回のデータセットの解析について考えてみると、モデル式は次のように書くことができるだろう。このモデル式は「$Group$ は Y を説明する」という意味をもっているが、より正確にいえば「$Group$ の水準ごとの平均値を当てはめたとき、Y の変動を有意に説明できるか」ということである。このようにモデル式を書くことは、自分が問題にしていることを明らかにするという意味で大変重要である。

$$Y = Group \quad (Group \text{ はカテゴリカル型})$$

6.2.2　平方和分解

さて、分散分析とは簡単にいえば変動の分解という作業を行っている。変動には全変動、群間変動、群内変動の 3 つを考えることになるが、これについてはデータをプロットして考えるとよい。したがって、まずは横軸にデータ番号を、縦軸にデータの値をとって散布図を描いてみるとよい。そこに全平均と群平均をそれぞれ当てはめてみるようなグラフを作成してみる（図 6.2）。

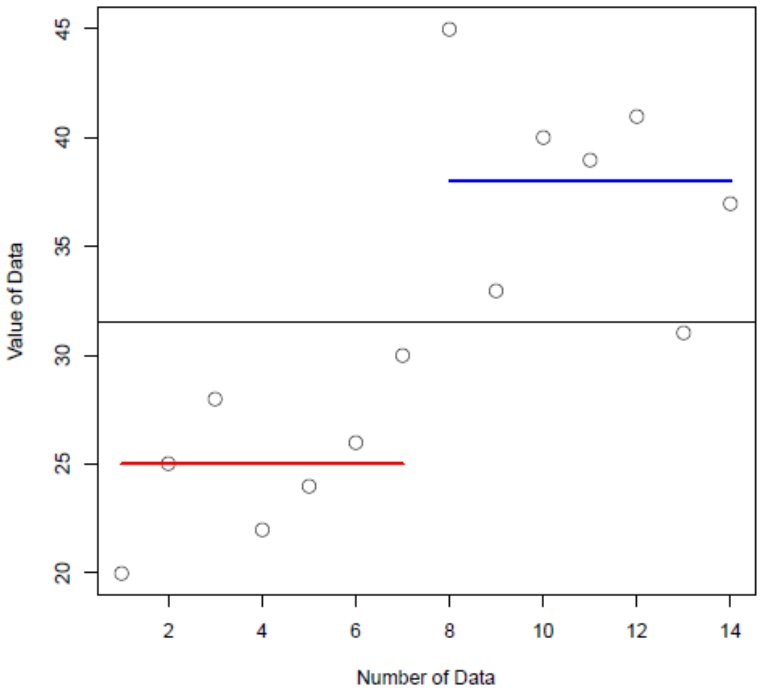


図 6.2　データに全平均と群平均を当てはめる

図 6.2 において、縦軸の中央辺りに引かれている線分が全平均であり、赤い線分が A 群の平均、青い線が B 群の平均である。分散分析はこの平均まわりのバラつき（分散）を分析する方法なので分散分析というわけである。ここで各群の群平均まわりのバラつきのことを**郡内変動**という（群内でのバラつきという意味である）。そして、その変動の大きさは**平方和**（SS, Sum of Squares）によって表される。分散の章で説明したように、平方和は分散を計算するための公式の分子である。

この平方和は値が大きいほどバラつきが大きいことを意味しているが、データ数が増えるに伴って増加するという性質をもっている。したがって、比較する群のサンプルサイズが等しい場合は単に平方和を比較しても問題なさそうであるが、そうでない場合は比較しても意味がないだろう。実際、今回のデータセットではサンプルサイズが同じなので、平方和を比較した場合、単純に平方和の値が大きい方が郡内変動が大きいことを意味する。

それでは実際に全変動 SSY、群間変動 SSG、群内変動 SSE をがどういうものであるかを説明しながら計算してみよう。まず SSY は全平均まわりのバラつきのことであり、全平均と原データの偏差の平方和である。図 6.2 でいえば、黒い線分（全平均）とデータポイント（原データ）の偏差の平方和である。より分かりやすくしたものが図 6.3 である。

$$SSY = 799.5$$

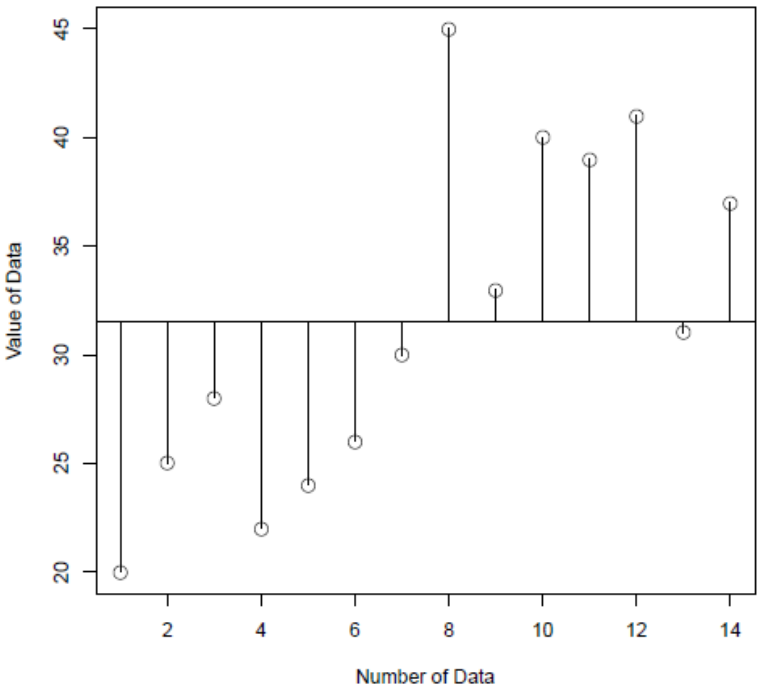


図 6.3　全平均と原データの偏差（全変動）

続いて SSG は全平均と群平均の偏差の平方和である。図 6.2 でいえば、黒い線と赤い線の偏差の平方和、および黒い線と青い線との偏差の平方和である。より分かりやすくしたものが図 6.4 である。

$$SSG = 591.5$$

最後に SSE は群平均と原データの偏差の平方和である。図 6.2 でいえば、赤い線と A 群のデータポイント、青い線と B 群のデータポイントの偏差の平方和である（図 6.5）。

$$SSE = 208.0$$

ここで SSY は以下のように分解されていることが分かるだろう（これを**平方和分解**という）。

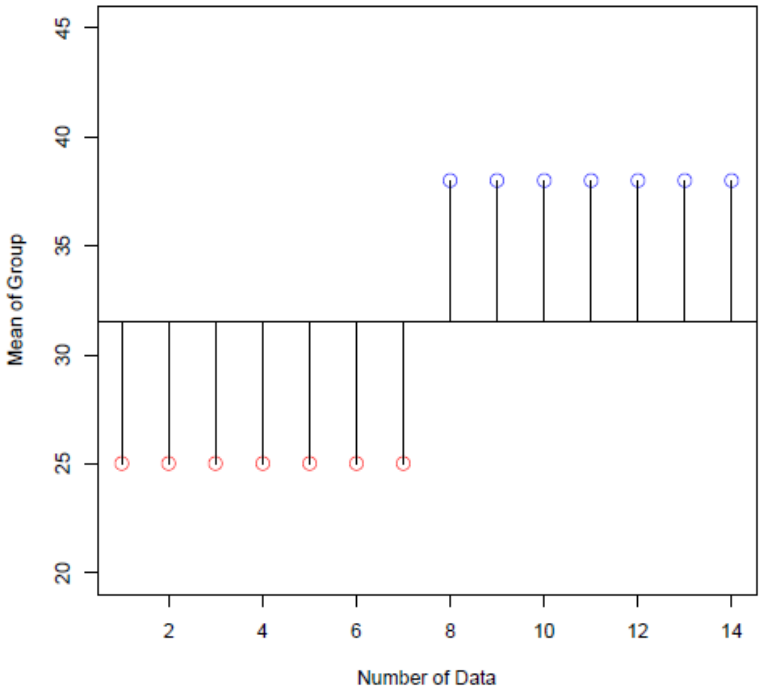


図 6.4　全平均と群平均の偏差（群間変動）

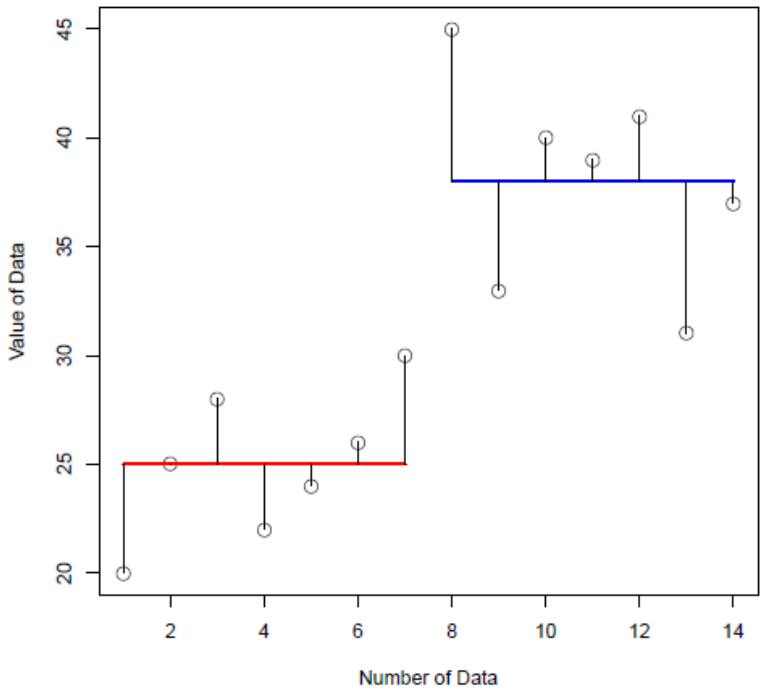


図 6.5　群平均と原データの偏差（群内変動）

$$SSY = SSG + SSE$$
$$799.5 = 591.5 + 208.0$$

この SSG と SSE を比較すれば、モデル式が意味するところの「$Group$ の水準ごとの平均値を当てはめることによって、Y の変動を有意に説明できるか」ということに対しての答えが得られるだろう。先に結論をいってしまえば、SSG（群間変動）がより大きく、SSE（群内変動）がより小さいほど Y の変動を有意に説明できるということになる。これについて、極端なデータセットの例をあげて説明してみよう。

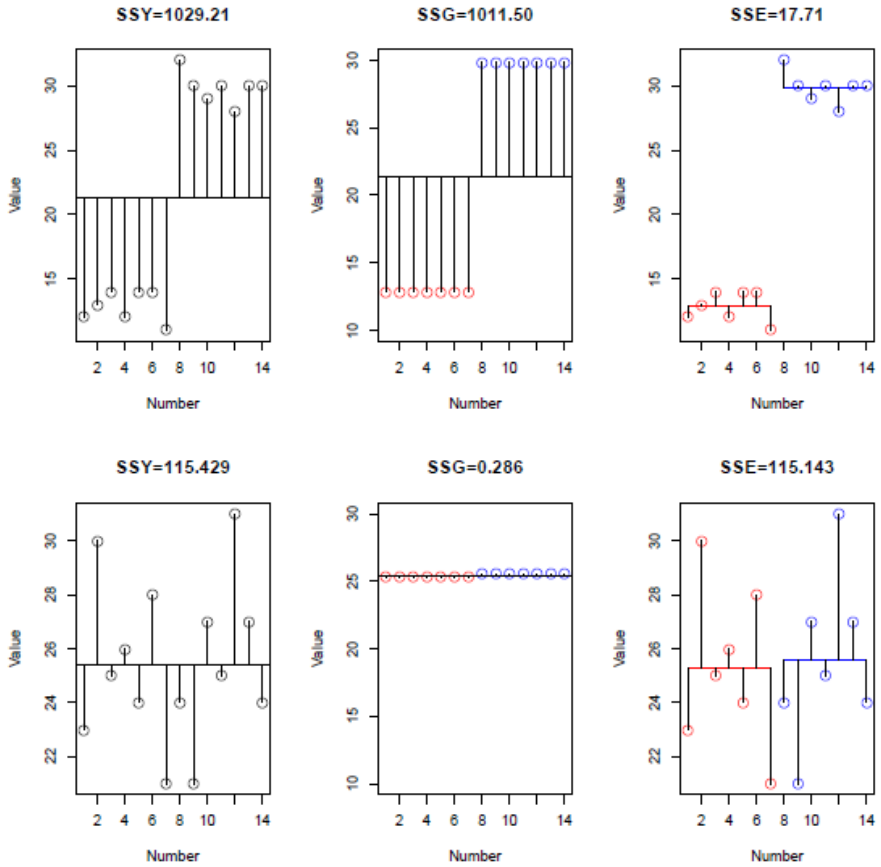


図 6.6　2 つの極端なデータセットの例

図 6.6 は左から順に全変動、群間変動、郡内変動を表した図になっており、上の 3 つは群間変動が大きく、群内変動が小さい場合の例である（有意な差が認められるデータセット）。一方で下の 3 つは群間変動が小さく、群内変動が大きい場合の例である（有意な差が認められないデータセット）。

有意な差が認められるデータセットでは、群平均を当てはめたときにデータの変動をうまく説明できている（SSE=17.71 となっている図）。それに対して有意な差が認められないデータセットでは、群平均を当てはめてもデータの変動を説明できておらず（SSE=115.143 の図）、全平均を当てはめた場合とさして変わらないのである。つまり、説明変数が応答変数

の変動を有意に説明するとは、群平均を当てはめたときに、群平均とデータの偏差が小さくなるということなのである。

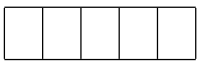
また、変動の分解という観点から見てみると、有意な差が認められるデータセットでは群間変動の平方和（SSG=1011.50）は全変動の平方和（SSY=1029.21）のほとんどを占めており、群内変動（SSE=17.71）が占める割合はわずかなものである。それに対して有意な差が認められないデータセットでは全く逆で、群内変動（SSE=115.143）が全変動のほぼ全てを占めており、群間変動（SSG=0.286）が占める割合はわずかである。

まとめると以下のようになる。

- 説明変数が応答変数の変動を有意に説明するとは、
- 群平均を当てはめたときに、群平均の直線がデータにうまくフィットしているということである。
- 換言すると、群平均とデータとの偏差（群内変動）が小さく、全平均と群平均の偏差（群間変動）が大きいということである。
- そして、このような状態のときに平均値が有意に異なるという結論が得られる。

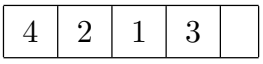
6.2.3 自由度

例えば、簡単なパズルゲームで自由度について考えてみる。以下の 5 つの箱にはそれぞれ 1 つの数字を入れることができる。全ての箱に数字を入れたとき、その合計が 10 になるように数字を入れる。最初の段階では 5 つの箱のいずれにも自由な数字を入れることができる。



仮にここで 4 という数字を箱に入れたとする。

まだ残りの 4 つの箱に自由な数字を入れることができる。そこで 2 つ目の箱に 2 を、3 つ目の箱に 1 を、そして 4 つ目の箱に 3 を入れてみたとしよう。



すると、ここで $4 + 2 + 1 + 3 = 10$ となるので、もはや残りの箱に入れられる数は限られている（5 つ目の箱に入れられる数字は 0 でなければならない）。すなわち、与えられた箱の数が 5 ならば、自由に数字を決めて数字を入れられる箱の数は 5 - 1 = 4 となる。これが自由度のイメージ的な理解の仕方である。

さて、変動の分解：

$$SSY = SSG + SSE$$

において、SSY を求める際に 14 個の偏差を計算することになるが、偏差の和は 0 になるので、13 個の偏差を計算すれば 14 個目の偏差はおのずと決まることになる（上の例において 4 つの箱に数字を入れた状態と同じである）。つまり、ここでの自由度は 13 であることがわかるだろう。次に SSG を求めるとき、全平均と 2 つの群平均との偏差は 2 つであるが、これもまた偏差の和が 0 になることから、どちらか 1 つの偏差が決まればもう一方の偏差が決定する。すなわち、SSG の自由度は 1 である。最後に SSE について、各群のサンプルサイズはそれぞれ 7 なので、群平均とデータの偏差はそれぞれ 7 つ求めることになる。つまり、各群の自由度は 7 - 1 = 6 となるので、SSE の自由度は 6 * 2 = 12 となる。

$$SSY \text{ の自由度 } (DFY) = 14 - 1 = 13$$
$$SSG \text{ の自由度 } (DFG) = 2 - 1 = 1$$
$$SSE \text{ の自由度 } (DFE) = (7 - 1) + (7 - 1) = 12$$

6.2.4　平均平方

先にも述べたように、平方和（SS, Sum of Squres）はデータ数の増加に伴って大きくなる。したがって、平方和自体を比較しても意味がないことが多い。そこで 1 自由度あたりの変動を考えるために、SS を DF で割ってやればよい。これはまさに分散そのものである。分散は平方和を自由度で割ることで求められるのであった（分散の章を参照のこと）。つまり、SSY、SSG、SSE それぞれの値を、それぞれの自由度で割ってやればよいのである。これを分散分析においては平均平方（MS, Mean of Square）と呼んでいる。まさに分散分析という名称にふさわしい内容であることが分かるだろう。

$$MSG = SSG/DFG = 591.5/1 = 591.5$$
$$MSE = SSE/DFE = 208.0/12 = 17.3$$
$$MSY = SSY/DFY = 799.5/13 = 61.5$$

6.2.5　F 値

説明変数（$Group$）が応答変数（Y）の変動を有意に説明しているかどうかは F 値によって判断される（直接的には p 値を見て判断する）。F 値は次式によって計算される。

$$F = \frac{MSG}{MSE} \tag{6.4}$$

図を用いた説明を上の方でしたが、そこでも書いたように有意な差が認められるデータとは群間変動が大きく、群内変動が小さくなるようなデータである。つまり、この F 値において MSG が大きく MSE が小さいほど F 値は高くなり、p 値は小さな値となる。

6.3 回帰分析の原理

6.3.1 直線の当てはめ

回帰分析モデル (重回帰分析モデル) とは、説明変数がいずれも連続型である場合のモデルのことをいうのであった。最も簡単なのは説明変数が 1 つだけである単回帰分析モデルである。したがって、ここでは単回帰分析について、少しだけ理論的な話を展開していくことにしよう。

まず 2 つの連続型の変数があったとき、それをグラフで表現するとなると散布図が適切な方法であろう。仮に $Y = \{12, 20, 22, 17, 15, 16\}$ というデータと、$X = \{7, 16, 20, 11, 12, 11\}$ というデータが得られているとしよう。これについて、横軸に X を、縦軸に Y をとってプロットしてみると図 6.7 のようになる。

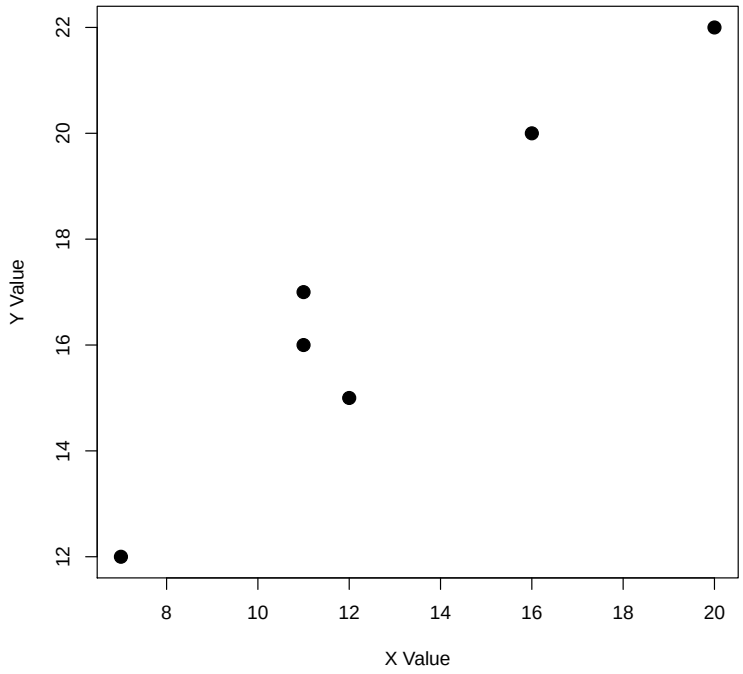


図 6.7 X と Y の散布図

ここで描かれたデータポイント (黒丸) に「最もよく当てはまる直線」を当てはめることを考えてみる。これは換言すると「ある直線を当てはめたとき、実測値とのズレが最も小さくなるような直線を当てはめる」ということである。このような直線の方程式、すなわち

$Y = \alpha + \beta X$ の α と β といったパラメータを推定する方法を**最小自乗法**という。

これについて理論的に学ぶことを望むのであれば、文献 [16] などを参考にすると良いだろう。ここでは数式による説明ではなく、どのようにして最適な直線が選ばれるかを図を用いて説明することにする (最も当てはまりの良い直線が選ばれる原理をイメージとしてつかみ取ってほしい)。

6.3.2　最適な直線の選択

手順 1　Y の平均を当てはめる

最初に行うべき作業は Y の平均を当てはめることである。図 6.7 に変数 Y の平均値 (17.0) を当てはめたものが図 6.8 に示してある。また、当てはめられた Y の平均値からのズレ (これを**残差**という) を点線で表してある。要はこの残差を最小にするような直線を選択しようというのが最小自乗法のアイディアである。

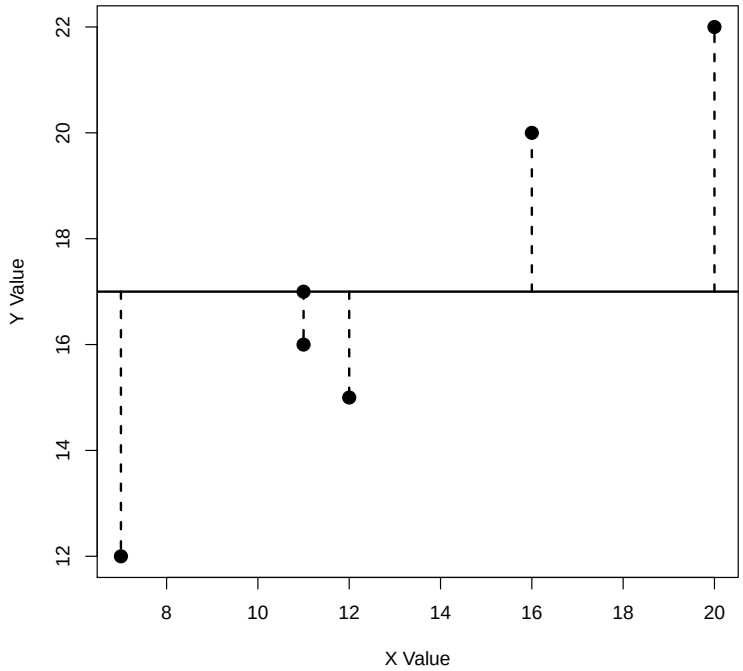


図 6.8　散布図に Y の平均を当てはめる

手順 2　X の平均を当てはめる

続いて X の平均を当てはめるという作業を行う (図 6.9)。図 6.9 において、先ほど当てはめた Y 平均の直線と、今回新たに当てはめた X 平均の直線が直交する点をバツ印として示しているのが分かるだろう。実はこの点を中心に Y 平均の直線を回転させていき、残差が最も小さいくなるような直線を決定するのである。

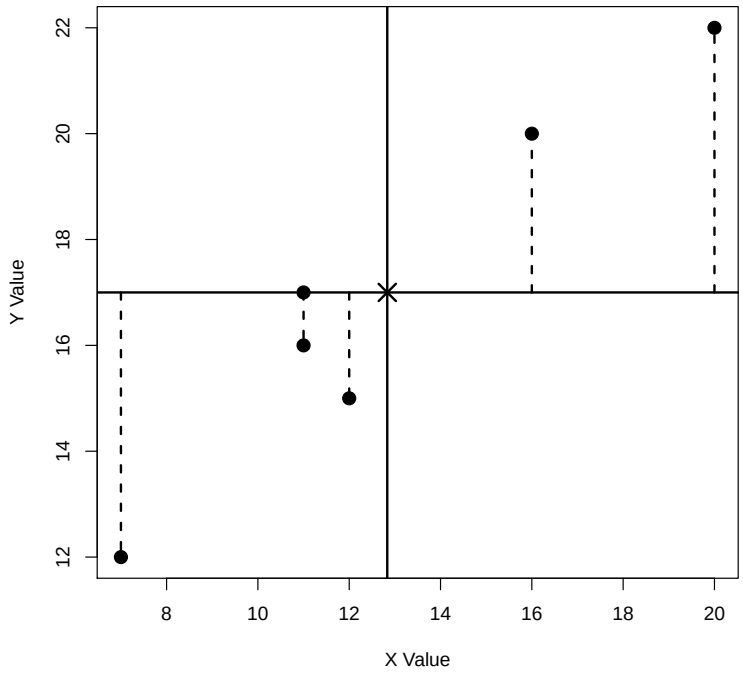


図 6.9 散布図に X の平均を当てはめる

手順 3 直線を回転させる

図 6.10 には最適な直線が当てはめられている。この回帰直線の方程式は $Y = 7.27 + 0.76X$ である。イメージとしては、図 6.10 の点線 (これは Y 平均である) を Y 平均と X 平均が直交する点 (今度はプラス符号をデータポイントとしてプロットしてある) を固定して、反時計回りに徐々に回転させていく様子を思い浮かべるとよいだろう。

以上が最小自乗法を用いて、最も当てはまりのよい (残差が最小になる) 回帰直線を決定するという作業のイメージ的な説明である。ただし、イメージ的とはいっても、基本原理はまさにこの通りのことをしているのである。

6.4 検定と推定

6.4.1 95% 信頼区間

従来的に心理学の分野では検定が幅を利かせてきたが、**信頼区間**を求めることにも大きな意味があることを知ってもらいたい。例えば、1 標本のデータについて考えてみよう。ここに $DAT = \{50, 44, 49, 61, 58, 55, 56\}$ というデータセットが用意されていたとする。このデータセットの平均値は 53.29 である。

それでは、この標本平均値は「母平均は 50 である」ということに等しいかどうかを考えて

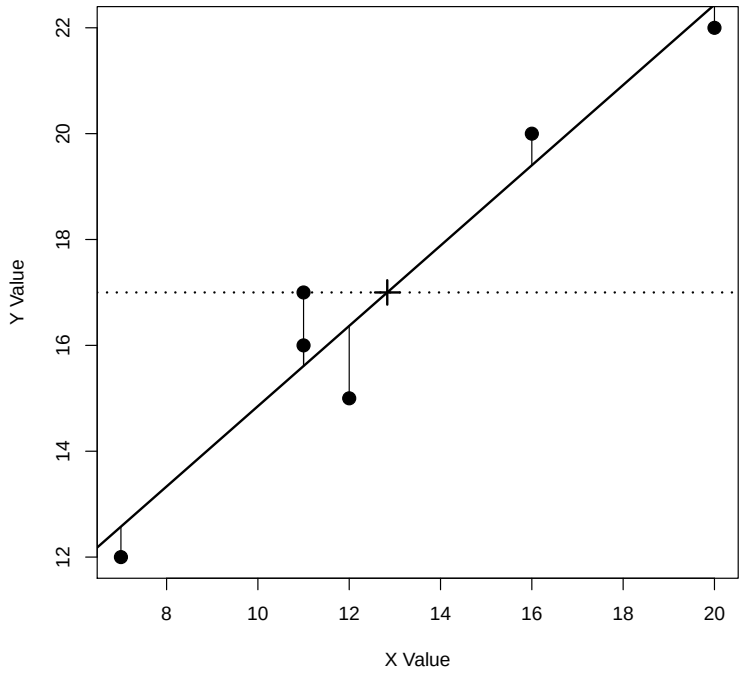


図 6.10　Y 平均と X 平均の直交点を中心に直線を回転させる

みよう。これは「母平均は 50 である」という帰無仮説について検定を行うことになる。実際に計算をしてみると、$t = 1.4785 (p = 0.1898)$ となるので、5% 水準の下では帰無仮説を採択することになる。

これとは別のアプローチで母平均 μ の 95% 信頼区間を求めて帰無仮説について考察することができる。仮に得られているデータが非常に多く (大きな標本の下では)、標本標準偏差 (s) が母標準偏差 (σ) に近似すると考えられるので、次式によって 95% の信頼区間を求めることができる (次式における $\sigma/\sqrt{n}$ は標準誤差である)。

$$\bar{y} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}$$

ただし、小さな標本の下では標本分散が母分散への近似が良くないため、1.96 の代わりに t 値を用いた次式によって計算されるべきである。

$$\bar{y} - t_{crit}\frac{s}{\sqrt{n}} < \mu < \bar{y} + t_{crit}\frac{s}{\sqrt{n}}$$

ここで t_{crit} は任意の自由度のときの**臨界 t 値**で、例えば、$df = n - 1 = 7 - 1 = 6$ のときの臨界 t 値は 2.447 である (これは R の `qt(0.025, 6)` もしくは `qt(0.975, 6)` とコマンドすることで得られる)。

実際に先ほどのデータセットから、母平均の 95% 信頼区間を計算してみると次のようになる。

$$53.29 - 2.447\frac{5.88}{\sqrt{7}} < \mu < 53.29 + 2.447\frac{5.88}{\sqrt{7}}$$

これを R で行えば次のようになる。

```
> dat <- c(50, 44, 49, 61, 58, 55, 56)    # データベクトル
> me <- mean(dat)                          # 平均値
> me
[1] 53.28571
> se <- sd(dat) / sqrt(length(dat))        # 標準誤差
> se
[1] 2.222336
> dat.t <- qt(0.975, 6)                    # t 値
> dat.t
[1] 2.446912
> me - dat.t * se   # 下側値
[1] 47.84785
> me + dat.t * se   # 上側値
[1] 58.72357
```

さて、以上のようにして求められた信頼区間 ($95\%CI = [47.85, 58.72]$) は統計学の教科書に載っている一般的なものであるが、より正確にいえば、これは**フィッシャーのフィデューシャル区間** (Fisher's diducial interval) と呼ばれるものである。この区間に 50 が含まれている場合、5% 水準の下で帰無仮説を採択することになる。逆にいうと、もしこの区間に 50 という値が含まれないならば、5% 水準で帰無仮説を棄却することになる。

これは回帰分析における、「推定された回帰係数は 0 である」という帰無仮説についての検定でも同様に用いることができる。例えば、R に最初から用意されているデータセット *cars* を用いて回帰分析を行った結果を以下に示す (このデータセットには自動車の走行スピード *speed* と制動距離 *dist* が含まれており、サンプルサイズは 50 である)。

```
> data(cars)          # cars を呼び出す
> names(cars)         # cars に含まれる変数を確認
[1] "speed" "dist"
> attach(cars)        # パスを通す
```

```
> result <- lm(dist ~ speed) # dist を応答変数、speed を説明変数とする
> summary(result)            # 係数表の出力

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
```

定数項、および係数の自由度は $df = n - p - 1$ である (n はサンプルサイズ、p は説明変数の数である)。したがって、自由度 $50 - 1 - 1 = 48$ の臨界 t 値は 2.01 であるから、定数項と係数それぞれの 95% 信頼区間は以下のようにして計算される。

```
> # 定数項の 95% 信頼区間を求める
> -17.58 - 2.01 * 6.76    # 下側境界値
[1] -31.1676
> -17.58 + 2.01 * 6.76    # 上側境界値
[1] -3.9924

> # 係数の 95% 信頼区間を求める
> 3.93 - 2.01 * 0.42      # 下側境界値
[1] 3.0858
> 3.93 + 2.01 * 0.42      # 上側境界値
[1] 4.7742
```

結局、定数項の 95% 信頼区間は $[-31.17, -3.99]$ となり、係数の信頼区間は $[3.09, 4.77]$ となった。両者どちらも、この区間内に 0 が含まれていないので、5% 水準の下では「推定されたパラメータは 0 である」という帰無仮説は棄却されることになる。

ブートストラップ

信頼区間を求める別の方法として**ブートストラップ法**がある。これは非常にユニークな方法で、得られている標本からデータをリサンプリング (繰返し抽出) して別の新たな標本を作り出す。このようにして得られた標本をブートストラップ標本というが、このブートストラップ標本を何個も用意し、各ブートストラップ標本から得た平均値をヒストグラムで表現してみると、それは正規分布の形状を描くようになる (これは中心極限定理と呼ばれる現象

である)。これにより、95% 信頼区間を求めようという発想である。

例えば、ここに 20 個のデータセットがあるとしよう。このようなデータセットは R の rnorm() という関数を用いると、簡単に用意することができる。rnorm() はある平均値と分散 (標準偏差) の正規分布に従うようなデータを任意の数だけ発生させることができる関数である。ただし、これは実行するたびに異なるデータが生成されるので、これ以降に載せる実行結果と読者諸君の実行結果とは異なることになる。

まず 20 個のデータを発生させる。これは平均が 50、標準偏差が 10 である正規分布 (母集団) から無作為抽出されたデータである。そしてこれを「元の標本」と呼ぶことにする。

```
> dat <- rnorm(20, mean=50, sd=10)
> dat
 [1] 40.48356 68.61982 44.63738 57.93450 35.11300 51.45787
 [7] 55.90245 48.52847 45.52874 53.97838 52.74332 51.09932
[13] 42.62078 63.12463 25.61103 57.57071 48.10338 48.22149
[19] 47.70679 49.29446
```

次に元の標本と同じ大きさのブートストラップ標本を作成する。今回は元の標本から、元の標本と同じ大きさ ($n = 20$) のデータをリサンプリングすることにしよう。そして、このブートストラップ標本を 10000 個用意し、それぞれのブートストラップ標本について平均値を求める。

このような手順を 1 回 1 回、手作業で行っていくと次のようになるが・・・

```
> bs1 <- sample(dat, 20, replace=TRUE)  # 1 つ目のブートストラップ標本
> bs2 <- sample(dat, 20, replace=TRUE)  # 2 つ目のブートストラップ標本
> bs3 <- sample(dat, 20, replace=TRUE)  # 3 つ目のブートストラップ標本
(途中省略)
> bs9999 <- sample(dat, 20, replace=TRUE)   # 9999 つ目のブートストラップ標本
> bs10000 <- sample(dat, 20, replace=TRUE)  # 10000 つ目のブートストラップ標本

> # 10000 個のブートストラップ標本の平均値を代入する変数を用意する
> bs.mean <- numeric(10000)
> bs.mean[1] <- mean(bs1)   # 1 つ目のブートストラップ標本の平均値
```

```
> bs.mean[2] <- mean(bs2)    # 2 つ目のブートストラップ標本の平均値
（途中省略）
> bs.mean[10000] <- mean(bs10000) # 10000 つ目のブートストラップ標本の平均値
```

これはまさに時間と手間の無駄というものである。こういう繰返し処理はコンピュータの得意技で、for() という関数 (プログラミングでは for 構文という) を使えばよいのである。

```
> a <- numeric(10000)
> for(i in 1:10000) a[i] <- mean(sample(dat, 20, replace=TRUE))
```

なお、ここで用いた sample() という関数は無作為抽出を行うための関数である。第 1 引数の dat と指定したものは、抽出の対象となるデータベクトルである。第 2 引数の 20 は「20 個サンプリングしなさい」と命令するためのもので、第 3 引数の replace=TRUE は復元抽出 (重複してサンプルすることを許すこと) を指定している。この第 3 引数を省略すると、非復元抽出 (重複を許さないこと) になるので注意しなければならない。

最後にヒストグラムを描き (hist() を使う)、95% 信頼区間を求めてみる (quantile() を使う)。10000 個のブートストラップ標本より得られた平均値のヒストグラムは図 6.11 のようになる。図中の点線は 95% 信頼区間の下側境界値と上側境界値を表している。

```
> hist(a)
> quantile(a, c(0.025, 0.975))
    2.5%    97.5%
45.11073 53.34062
> abline(v=c(45.11, 53.34), lty=3, lwd=2)
```

ところで、このブートストラップについて詳しく知りたいのであれば、文献 [13] が日本語で書かれた本の中では分かりやすい。英語の文献を読むことに抵抗がなければ、文献 [14] や文献 [3] などが薦められる。

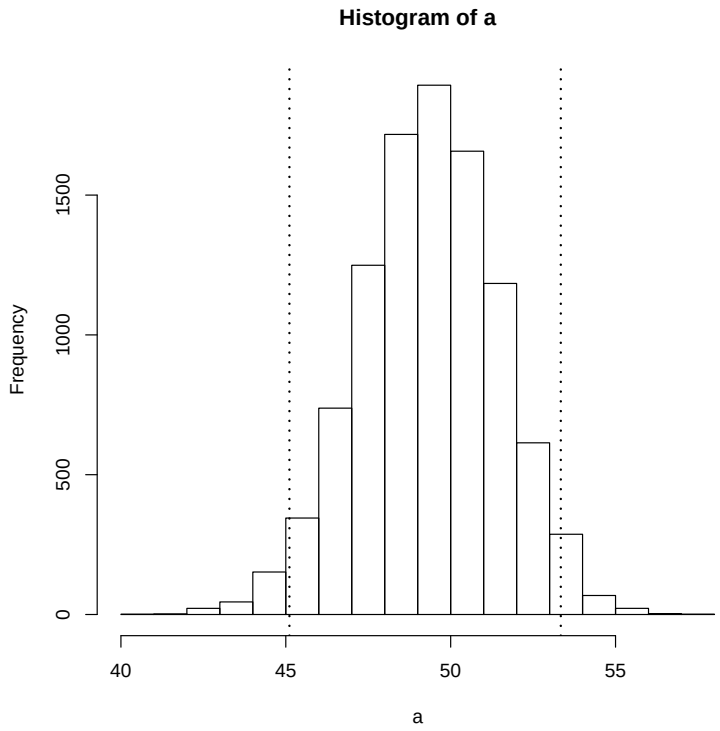


図 6.11　10000 個のブートストラップ標本の平均値ヒストグラム

参考文献

[1] 足立浩平. 多変量データ解析法　心理・教育・社会系のための入門. ナカニシヤ出版, 2006.

[2] Alan Agresti. *An Introduction to Categorical Data Analysis*. Jhon Wiley and Sonx, Inc., 1996. (渡邉ほか 訳. カテゴリカルデータ解析入門. サイエンティスト社, 2005.).

[3] A. C. Davison and D. V. hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997.

[4] Annette J. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, 2002. (田中ほか 訳. 一般化線形モデル入門. 共立出版, 2008.).

[5] Brian S. Everitt. *An R and S-PLUS Companion to Multivariate Analysis*. Springer, 2005. (石田基宏 訳. R と S-PLUS による多変量解析. シュプリンガー・ジャパン, 2007.).

[6] 藤越康祝, 菅民郎, 土方裕子. 経時データ分析. オーム社, 2008.

[7] 藤越康祝, 杉山高一, 狩野裕. 経時データ解析の数理. 朝倉書店, 2009.

[8] Alan Grafen and Rosie Hails. *Modern Statistics for the Life Sciences -Learn to analyse your own data-*. OXFORD UNIVERSITY PRESS, 2008. (野間口謙太郎, 野間口眞太郎 訳. 般線形モデルによる生物科学のための現代統計学　－あなたの実験をどのように解析するか－. 共立出版, 2007.).

[9] 一石賢. 道具としての統計解析. 日本実業出版社, 2004.

[10] 稲垣宣生. 数学シリーズ　数理統計学. 裳華房, 2004.

[11] 岩崎学, 中西寛子, 時岡規夫. 実用統計用語辞典. オーム社, 2004.

[12] Jhon M. Chambers and Trevor J. Hastie. *Statistical Models in S*. AT&T Bell Laboratories, 1992. (柴田里程 訳. S と統計モデル－データ科学の新しい波－. 共立出版, 1999.).

[13] 小西貞則, 越智義道, 大森裕浩. 計算統計学の方法　ブートストラップ・EM アルゴリズム・MCMC. 朝倉書店, 2008.

[14] Bryan F. J. Manly. *Randomization, Bootstrap And Monte Carlo Methods in Biology*. Chapman & Hall, 2006.

[15] 間瀬茂, 神保雅一, 鎌倉稔成, 金藤浩司. 工学のためのデータサイエンス入門－フリーな統計環境 R を用いたデータ解析－. 数理工学社, 2004.

[16] Michael J. Crawley. *Statistics: An Introduction useing R.* John Wiley & Sons, Ltd, 2005. (野間口謙太郎, 菊池泰樹 訳. 統計学: R を用いた入門書. 共立出版, 2008.).

[17] 水野欽司. 多変量データ解析講義. 朝倉書店, 2002.

[18] 森敏昭, 吉田寿夫. 心理学のためのデータ解析テクニカルブック. 北大路書房, 2006.

[19] 豊田秀樹. 違いを見ぬく統計学　実験計画と分散分析入門. 講談社ブルーバックス, 2006.

[20] 豊田秀樹. 共分散構造分析　Amos 編. 東京図書, 2007.

索引