質的なものである．**量的データ**（Quantitative data）には二種類あり，一つは身長，体重，血圧などの測定できる連続変数である．もう一つは一家族当たりの子供の数，月当たり子供当たりの喘息発作回数など，離散変数である．したがって，頻度データ（count data）というのは離散で連続なデータということになる．連続変数は正規分布のことが多いが，非正規分布のこともある．正規分布ということは，データをヒストグラムに描くと独特の『ベル型（bell-shaped）』になるということである．実際には，データが一つの中心点の周囲に集まり，分布がこの点の回りで対称であれば，通常は正規分布に非常に近いと考えてよい．その場合は，正規性を仮定する多くの検定が可能となる．また，平均値と中央値は近い値をとると予想できる．非正規分布は非対称の（歪んだ）分布となり，平均値と中央値は異なる．非正規分布変数の例としては，ある集団の年齢や収入がある．時に非対称の原因が，実はデータの誤りである外れ値による場合があるので，注意して調べる必要がある．

『非正規分布（non-Normally）』データではなく『ノンパラメトリック（non-parametric）』データと呼ぶことがあるが，これは誤った名称であることを覚えておこう．パラメータとはモデルに伴うものであり，『ノンパラメトリック』データというとモデルに当てはまらないデータということになる．しかし後で見るように，これは統計手法の非常に狭義の定義である．量的データの重要な特徴は，実際に意味をもつ数値で扱える点である．例えばデータの平均値が計算できる．これは質的データとは対称的である．質的データでは多くの場合，数値そのものは都合のよいラベルにすぎない．

質的データ（Qualitative data）はカテゴリーになりがちである．例えば男性か女性，ヨーロッパ人かアメリカ人か日本人，病気か健康か，のようにである．これらは**名義**（nominal）あるいは**カテゴリー**（categorical）といわれる．もし２つのカテゴリーしかなければ**二値**（binary）データという．カテゴリーは順序のこともある．例えば，『良好（get better）』，

『不変 (stay the same)』, 『悪化 (get worse)』などである. これは**順序** (ordinal) データである. 多くの場合, これを例えば1, 2, 3のようにスコア化する. しかし, 2人の患者がいて, 1人が良くなりもう1人が悪くなったとしよう. このとき, 平均として変化はなかったと言っても意味がない (統計家とは頭をオーブンの中に入れ, 足を冷蔵庫の中に入れても, 平均として快適だという人である!). 順序データの重要な点は順序になっていることだが, 重みのつけ方には明白なものはない. 例えば, 『健康 (healthy)』, 『病気 (ill)』, 『死亡 (dead)』をどのように重み付けすればよいかは明らかではない (後で述べるように, 順序カテゴリーに連続する整数を割り当てスコア化し, それを連続変数のように扱うことが多い. あるいは, 変数を二つに分けて二値として扱うのが良いこともあろう). 数えられるデータ, 例えば一家族の子どもの数はどうだろうか. これは明らかに順序データであるが, ここで重要な点は計算可能なことである (一家族当たり子供2.4人には意味がある). これは, ときに**比** (ratio) としての特性をもつと表現される. 4人子供がいる家族は, 2人子供がいる家族の2倍の子供がいる. しかし, 例えば, 『強く支持する (strongly agree)』, 『支持する (agree)』, 『支持しない (disagree)』, 『全く支持しない (strongly disagree)』という4カテゴリーの順序変数があり, それぞれに1から4までスコア化したとしよう. スコア4『強く支持する』は, スコア2『支持しない』の2倍という言い方はできない.

連続データをカテゴリー化することで質的データを作ることができる. 血圧は連続変数だが, 『正常血圧 (normotension)』と『高血圧 (hypertension)』に分けることができる. 多くの場合, このほうが要約するのは簡単である. 例えば集団の10%が高血圧であると言ったほうが, 平均値と分散が示されるよりわかりやすい. もちろん, 後者からは前者 (さらにそれ以上) を導き出すことはできる.

従属変数が連続のときには, 第2章で取り上げる**重回帰** (multiple re-

gression）を使う．二値データのときは，第3章のロジスティック回帰や第4章の生存時間の解析を使う．もし従属変数が順序であれば第6章の順序回帰を使い，それが頻度データであれば第6章のポワッソン回帰を使う．一般的に言って，独立変数に関するデータの型はあまり重要ではない．

1.4 有意性検定

χ^2検定やt検定などの有意性検定とP値の解釈については，『初めて学ぶ医療統計学』[1]（Statistics at Square One）で述べた．統計学的有意性検定は，**帰無仮説**（null hypothesis）を立てることから始まる．そしてデータを収集する．帰無仮説を使って，観察されたデータが帰無仮説と一致するかどうか検定する．例えば，肥満患者を対象に新しい食事療法と標準的な食事療法で，体重減少を比較する臨床試験について考えてみよう．帰無仮説は，2種類の食事療法間で患者の体重変化について差が無いことである．結果は，2種類の食事療法後における平均体重変化の差である．2種類の食事療法間で差無しという帰無仮説が正しいとして，観察された平均の差（もしくはそれより極端な結果）が得られる確率を計算する．この確率（P値）が十分小さければ帰無仮説を棄却して，新しい食事療法と標準的なものは違うと考える．具体的には2つの食事療法群の体重変化の平均の差を，その推定標準誤差で割り，この比を（サンプル数が少ない場合は）t分布，（サンプル数が多い場合は）正規分布と比較するのが通常の方法である．

上で述べた方法はStudentのt検定として知られている．しかし，推定値をその標準誤差で割り，正規分布と比較するという検定の方式は**ワルト検定**（Wald test）として知られている．

実際，非常に多くの統計学的検定手法があるが，正規分布するデータでは通常同じP値を得る．しかし，違う型のデータでは異なる結果を

得ることになる．医学文献では，通常3種類の検定が使用されており，その構成と違いについて基礎的なことを知っておきたい．それらは**ワルト検定**（Wald test），**スコア検定**（score test），そして**尤度比検定**（likelihood ratio test）である．非正規分布データでは異なるP値になるが，データ数が増えるにしたがって一致してくる．この3つの検定の基礎は［付録2］に記載した．

1.5 信頼区間

統計学的検定の問題点は，P値がデータの大きさに依存することである．たとえ2つの治療間の差は小さくても十分な数のデータがあれば，ほとんどの場合有意差を証明できる．そこで，検定だけでなく平均効果の推定値に関する分析結果を示すのが大切になる．つまり推定値の精度，例えば信頼区間を示す．信頼区間を理解するために，母集団と標本の違いについて考えてみよう．母集団とはわれわれが一般化したいグループであり，例えば糖尿病患者や中年男性などである．母集団は**パラメータ**（parameters）をもっている．例えば，糖尿病患者の平均HbA_{1c}や中年男性の平均血圧などである．モデルは母集団をモデル化するために使われるので，モデルパラメータは母集団のパラメータということになる．われわれはこのモデルパラメータに関する**推定値**（estimates）を得るために，標本を取ってくることになる．真のモデルパラメータに等しい推定値を得ることはできないが，標本数が多くなれば真の値により近い推定値を得ることができる．推定値の信頼区間は，このことを数量化するのに役立つ．母集団平均の95%信頼区間とは，一定の大きさの標本を100回取りその都度平均と95%信頼区間を計算したとき，信頼区間の95回は真のモデルパラメータを含むと期待されるということである．しかし一般には，1つの標本から求めた95%信頼区間は，母集団パラメータを含んでいる可能性が95%だと理解する．

先に述べた食事療法を例にとると，信頼区間は新しい食事療法の効果をどのくらいの精度で推定したかを表す．もし新しい食事療法と古い食事療法に違いがなければ，信頼区間はゼロ（つまり差無し）を含むことになる．

1.6 モデルを使った統計検定

t 検定は，二群の連続変数の平均値を比較する際に使う．これは線形モデルで書くことができる．先の例では，2 つの食事療法のどちらであっても治療後の体重は連続変数である．ここで，主たる予測変数 x は食事療法であり，例えば標準食事療法が 0，新しい食事療法が 1 のように二値変数で表される．結果変数は体重であり，交絡変数はない．したがって当てはめるモデルは

$$\text{体重} = b_0 + b_1 \times \text{食事療法} + \text{残差}$$

となる．このモデルの FIT 部分は $b_0 + b_1 \times$食事療法であり，食事療法の効果推定値が分かれば，ある人の体重が食事療法後どれくらいになるかを予測できる．残差はほぼ正規分布すると仮定する．食事療法に関する係数 b_1 は平均がゼロの母集団から得られる，というのが帰無仮説になる．すなわち帰無仮説の下では，β_1 という母集団パラメータはゼロと仮定する．

モデルを使用することで，仮説をはっきりさせることができる．モデルのすぐれた特徴として，容易に拡張できる点がある．これは，検定とは対照的である．したがって，ベースラインである食事療法前の体重が（たまたま）二つの群で異なっていて，治療後の体重がそれと関連している場合でも，交絡変数としてそれを取り入れることができる．

これについて重回帰を用いる方法は，第 2 章で詳しく述べる．χ^2 検定のモデル扱いについては，第 3 章のロジスティック回帰で述べる．

1.7　モデルの適合度と解析：探索的解析と検証的解析

データ解析には二つの側面がある．それは探索的なものと検証的なものである．**検証的解析**（confirmatory analysis）では，あらかじめ立てておいた仮説を検定する．この場合，通常は有意性検定を行うことになる．臨床試験での治療効果の検定はその良い例である．**探索的解析**（exploratory analysis）では，データから何がいえるのかを探すことになる．例としては，コホート研究で危険因子を探る場合がある．そこでの知見は仮のものであり，続いて行う研究で検証する必要がある．また，P値は多分に装飾的な意味しかない．しばしばこの二つの解析を，同じ研究で行うことがある．例えば，ある臨床試験を解析する際，非常に多くの結果変数が測定されていたかもしれない．主たる結果としてプロトコル(計画書）に明記されたものは，検証的解析を行うことになる．しかし，副作用に関することなど大量の情報を解析することもできる．それらも報告すべきであるが，それらは解析して出てきた結果であり，あらかじめ決めておいた仮説から出てきたものではないことに注意してほしい．研究では，このような情報を無視しては論理的ではないだろう．また，全く予想していなかった有意な結果を無視してしまうのもつらいが，そうすべきなのである．

監査（audit）と**研究**（research）とを区別することが有用なこともあろう[訳注]．前者は多くの場合は記述的であり，ある特定の時間や場所の情報を提供する．一方，研究になると，他の時間や場所にも当てはまるよう一般化するものである．

1.8　コンピュータを駆使した方法

これからこの本で述べる多くの理論は，例えば正規分布などデータ分

訳注：監査とはいわゆる追試的確認結果であり，研究とは新規的な結果と思われる．

布が規定されていることが前提となる．モデルは使うが，あまり実際の分布には依存しないという現代的方法がある．それらはコンピュータへの依存度が大きく，最近まで簡単に利用できる方法ではなかった．しかし，それらの方法は徐々に広く使用されるようになってきている．完成されたものの一つとして，**ブートストラップ法**（bootstrap）については［付録 3］で述べた．

1.9 ベイズ流の方法

モデルに基づくアプローチをとると，次のような文章に行きつく．すなわち，『モデルMが与えられたとき，データDを得る確率はPである』（given model M, the probability of obtaining data D is P）となる．これは**頻度論的**（frequentist）アプローチといわれる．そこでは，母集団パラメータは固定であると仮定する．しかし，多くの研究者が言いたいことは，モデルMが正しい確率なのである．つまり，『データDが与えられたとき，モデルMが正しい確率はいくらか』（given the data D, what is the probability that model M is the correct one？）ということである．例えば，食事療法がうまくいく確率を知りたいのである．このような形で記載をすると，個々の患者が決断する際に非常に有用となる．これが『ベイズ流（Bayesian）』という考え方であり，そこでは母集団パラメータは個々人で異なっていてもよい．この本は大部分，頻度論的アプローチに基づいている．また，多くのコンピュータパッケージも然りである．さらに詳しいことは，第5章と［付録 4］に述べてある．

1.10 文献上の統計結果を報告するには

医学文献上の統計結果の記載は不十分なことが多い．ここでは一般に使えるポイントを簡単に紹介する．続く章では，特別な分析方法の記載

法について考えることにする．

もっと知りたい人は，Lang and Secic の本[4]がお勧めである．そこでは，医学文献で用いられる統計のさまざまな記載方法について書いてある．Altman らの本[3]には，統計解析を読んだり書いたりする際のチェックリストがある．臨床試験に関しては，CONSORT 声明[5]を参照してほしい．

- 被験者をどのように募集し，何人が試験に参加し，何人が脱落したかを必ず記載しなさい．臨床試験では，何人を試験に参加してもらうためにスクリーニングしたかを言及し，脱落者について治療群ごとに記載しなさい．
- 使用したモデルとその基礎になる仮定，そしてそれをどのように確認したかを記載しなさい．
- 主たる結果の推定値を記載しなさい．その際 P 値だけでなく，95% 信頼区間のような精度も示すべきである．また，意味のある推定値を表記することが重要である．臨床試験では，治療群ごとの効果の平均に興味がある対象かもしれないが，効果の主たる指標は差の平均と**差に関する**（for the difference）信頼区間である．それは多くの場合，治療群ごとの平均の信頼区間からは導くことはできない．
- P 値を得た方法（ワルト，尤度比，スコア）あるいは実際の検定法を記載しなさい．
- 時には二値（高血圧患者の割合など）でデータを**記載**（describe）することは有用である．しかし，**解析する**（analyse）ときは連続値（血圧など）で行いなさい．
- 使用した統計ソフトを記載しなさい．何か特殊な検定手法を用いたときなどは，その理由がわかるだろう．『自家製（home grown)』プログラムの結果は，その信頼性についてさらに確認が求められるかもしれない．

1.11 文献上の統計の読み方

- どんな母集団から得られたデータか？　結果は一般化できるものか？　欠損値は多くないか？　多くの人が協力を拒否していないか？
- 解析は検証的なものか探索的なものか？　研究か監査か？
- 正しい統計モデルが使用されているか？
- 『有意な効果がみられた（a significant effect was found）』などの記載に満足してはいけない．効果の大きさはどれくらいか，それが患者にとって意味があるかを確認しなさい（これはしばしば『臨床的に有意な効果（clinically significant effect）』と記載する）．
- 結果はモデルに関する仮定にかなり依存しているか？　多くの場合，結果はモデルに対して非常に『安定したもの（robust）』であるが，検討しておく必要はある．

[多肢選択問題]

1. データの型

乳がん患者の調査を行った．次に挙げるデータは，カテゴリー，二値，順序，量的連続，量的離散（頻度データ）のいずれか，示しなさい．

（i）治療を受けた病院
（ii）患者の年齢（年の単位で）
（iii）術式
（iv）乳がんのグレード
（v）運動後の心拍数
（vi）身長
（vii）就労状態（就労/非就労）
（viii）年当たり患者当たりの一般内科への受診回数

2. 原因・交絡・結果変数

正しいか間違っているかで答えなさい.

この章で述べた食事療法の試験について

(i) 結果変数は治療後の体重である.

(ii) 食事療法のタイプは交絡因子である.

(iii) 喫煙習慣は潜在的交絡因子である.

(iv) ベースラインである食事療法前の体重は入力変数になりうる.

(v) 食事療法は離散的な量的変数である.

3. 基礎統計

認知行動療法 (CBT) と薬物療法を比較するために試験を実施し, 次の結果を得た. 6ヵ月後のうつ指数の平均, CBT=5.0, 薬物療法=6.1, 差=1.1, P値=0.45, 95% 信頼区間=−5.0〜6.2であった.

(i) CBTは薬物療法と同等である.

(ii) P値を求めるのに適切な方法はt検定である.

(iii) 試験の結果は有意ではない.

(iv) 45% の確率でCBTが薬物療法より優れている.

(v) 同じ状況で同じサイズの試験をもう一度実施したら, 平均の差が−5.0〜6.2になる確率は95% である.

[演習問題の解答]

1. データの型

(i) カテゴリー, (ii) 連続 (非正規), (iii) カテゴリー, (iv) 順序, (v) 連続, (vi) 連続, (vii) 二値, (viii) 量的離散

2. 原因・交絡・結果変数

(i) 正解, (ii) 誤り (食事が2種類あり, それらは原因なので), (iii) 正解, (iv) 正解, (v) 誤り (二値変数なので)

3. 基礎統計

（i）誤り（有意ではないが信頼区間は広いため），（ii）正解，（iii）正解，（iv）誤り（CBTと薬物治療が同等と仮定して，観察された差あるいはそれ以上の結果を得る可能性が45%ということなので），（v）誤り（新たな試験の信頼区間と前の試験の信頼区間は重なるかもしれないが，新たな試験での差の平均値は母集団平均値ではないから）

■文献

1 Swinscow TDV. *Statistics at Square One, 9th edn.* (revised by MJ Campbell). London: BMJ Books, 1996.

2 Chatfield C. *Problem Solving. A statistician's guide.* London: Chapman and Hall, 1995.

3 Altman DG, Machin D, Bryant TN, Gardner MJ eds. *Statistics with Confidence, 2nd edn.* London: BMJ Books, 2000.

4 Lang TA, Secic M. *How to Report Statistics in Medicine: annotated guidelines for authors, editors and reviewers.* Philadelphia, PA: American College of Physicians, 1997.

5 Begg CC, Cho M, Eastwood S, Horton R, Moher D, Olkin I *et al.* Improving the quality of reporting on randomised controlled trials: the CONSORT statement. *JAMA*; **276**:1996; 637–9.
（訳注：CONSORT声明の改訂版が2001年に出た．その日本語訳は医学のあゆみ2002；201(10)：790–798，862–867．に見られる．）

2 線形重回帰

要 旨

連続な結果変数をモデル化する分析方法として，多くの場合**線形重回帰** (multiple linear regression) が適切である．単回帰については Swinscow[1] ですでに述べられた．単回帰では入力変数は 1 つで連続であったが，重回帰では入力変数を 2 つ以上へと一般化するのである．入力変数として，連続変数やカテゴリー変数を扱うことも可能である．さらに，カテゴリー変数を扱うための**ダミー変数** (dummy) や**標識変数** (indicator variables) についても述べる．また，例えば**てこ比** (leverage) や**影響度** (influence) のような概念を用いれば，それぞれのデータ値のモデルへ与える感度を調べることもできる．重回帰とは，**分散分析** (analysis of variance) あるいは**共分散分析** (analysis of covariance) の一般化といえる．ここで使うモデル化の手法は後の章でも有用である．

2.1 モデル

重回帰の基本モデルは次のようなものである．

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i . \qquad (2.1)$$

誤差項 ε_i は平均が 0 で，標準偏差が σ の正規分布であると仮定する．

第 1 章でモデル構造について述べたように，ここでの関連は線形であり，誤差項は正規分布である．

ここで，y_i はユニットまたは被験者 i の出力変数である．そして k 個の入力変数，X_{i1}，$X_{i2}, \ldots, X_{ik}$ がある．通常 y_i は**従属変数** (dependent) と呼んでいる．一方，入力変数 X_{i1}，$X_{i2}, \ldots, X_{ik}$ は**独立変数** (independent

variables) と呼んでいる．後者は連続変数でも名義変数でもよい．しかし，k 個の X は互いに独立である必要がないので，『独立』(independent) という言葉使いは誤っている．**説明変数** (explanatory variables) や**予測変数** (predictor variables) と言うこともある．入力変数は，**回帰係数** (regression coefficient) β_1, $\beta_2, \ldots, \beta_k$ と結び付いている．また，付加的な定数項 β_0 もある．これらは**モデルのパラメータ** (model parameters) である．

方程式 (2.1) の右辺の最初の部分は

$$\mathrm{LP}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

と書くことができる．LP_i は**線形予測子** (linear predictor) として知られており，入力変数によって予測される y_i の値である．差である $y_i - \mathrm{LP}_i = \varepsilon_i$ は，**誤差** (error) 項である．

この予測誤差の二乗和を最小にするようにして，モデルの推定値 $b_0, b_1, \ldots b_k$ を求める．このような推定値を**最小二乗** (ordinary least squares) 推定と呼ぶ．この推定値を使うことで，モデルを当てはめたときの値 y_i^{fit}，さらに第1章で述べたように測定された残差 $e_i = y_i - y_i^{fit}$ を計算することができる．ここで明らかなように，残差は誤差項を推定するのに用いている．詳細は Draper and Smith[2] を参考のこと．

2.2 重回帰の使い方

1．交絡の影響を考慮して，連続の出力変数に対する入力変数の効果を調整する．例えば，喫煙習慣を考慮に入れて，体重に対する食事療法の効果を調べる場合を考えてみよう．ここでは，従属変数は臨床試験から得られる結果である．独立変数は2つの治療集団 (0/1 の二値変数)，喫煙 (週当たりの箱数で連続変数)，そして始めの体重 (初期体重) である．重回帰を使えば，初期体重や喫煙習慣の違いを考慮に入れて治療群間の結果を比較できる．

2．与えられた入力変数に対して結果の値を予測する．例えば，ある年齢や身長の被験者の FEV_1 が予測できれば，測定された FEV_1 が予測値の何%であるかや，測定した FEV_1 が予測された値の80%を下回るかどうかなどを判断することができる．

3．カテゴリーの出力変数に対する効果を一度に分析する別の方法として**分散分析**（analysis of variance）があるが，重回帰でも同じ結果が得られる．

2.3　2つの独立変数

独立変数が2つの場合から始めることにする．この場合の独立変数は連続でも二値でもよい．3つの場合が考えられる．両方の変数が連続，両方が二値（0/1），そして一方が連続他方が二値の場合である．実際のデータを例に考えてみよう．

事　例

Swinscow[1] の中でも示された，15人の子供について，肺の解剖学的死

表2.1　15人の子供の肺機能検査結果

児番号	死腔 (mℓ)	身長 (cm)	喘息 (0=無,1=有)	年齢 (歳)	気管支炎 (0=無,1=有)
1	44	110	1	5	0
2	31	116	0	5	1
3	43	124	1	6	0
4	45	129	1	7	0
5	56	131	1	7	0
6	79	138	0	6	0
7	57	142	1	6	0
8	56	150	1	8	0
9	58	153	1	8	0
10	92	155	0	9	1
11	78	156	0	7	1
12	64	159	1	8	0
13	88	164	0	10	1
14	112	168	0	11	0
15	101	174	0	14	0

腔と身長に関するデータについて考えてみよう．仮に15人の子供のうちで8人が喘息，4人が気管支炎としよう．データは，**表2.1**のとおりである．

2.3.1 1つが連続，1つが二値の独立変数の場合

Swinscow[1]の中で提示された問題は，死腔と身長の間に関係があるかないかであった．ここでは，喘息児と非喘息児では死腔と身長の関係が違うのか，を調べてみる．

この場合，身長と喘息という2つの独立変数を扱うことになる．いくつかの考えられるモデルがある．

1. 平均は異なっているが，勾配と切片は2つのグループで同じ場合

モデルは，次のようになる

$$死腔=\beta_0+\beta_{身長}\times 身長 \qquad (2.2)$$

図2.1に示したように，Swinscow[1]の中の線形単回帰モデルである．

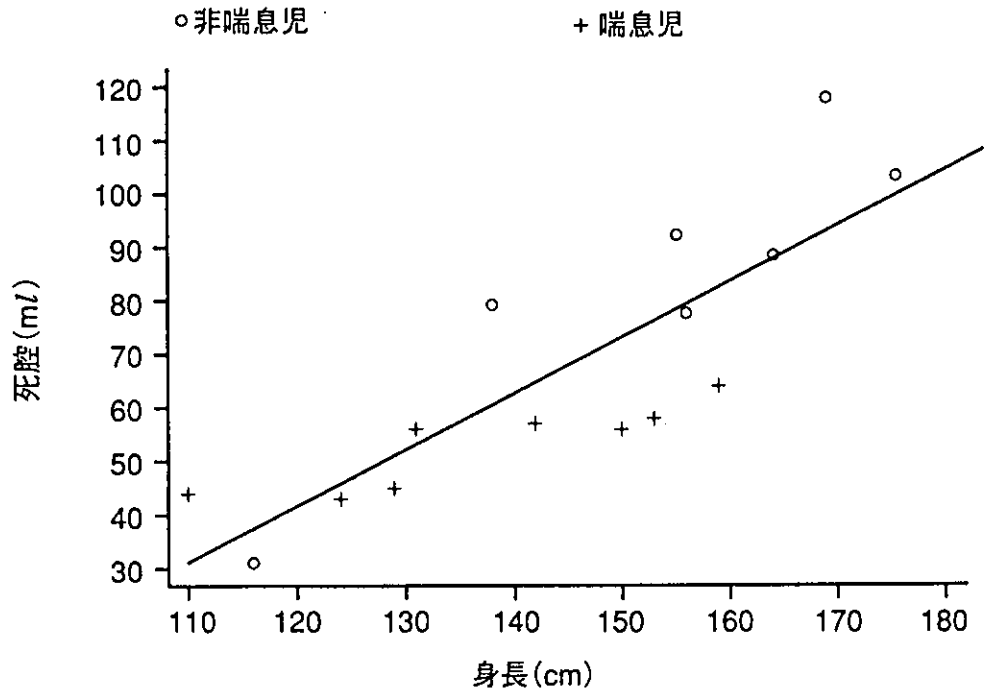

図2.1　喘息を考慮しない場合の死腔と身長との関係

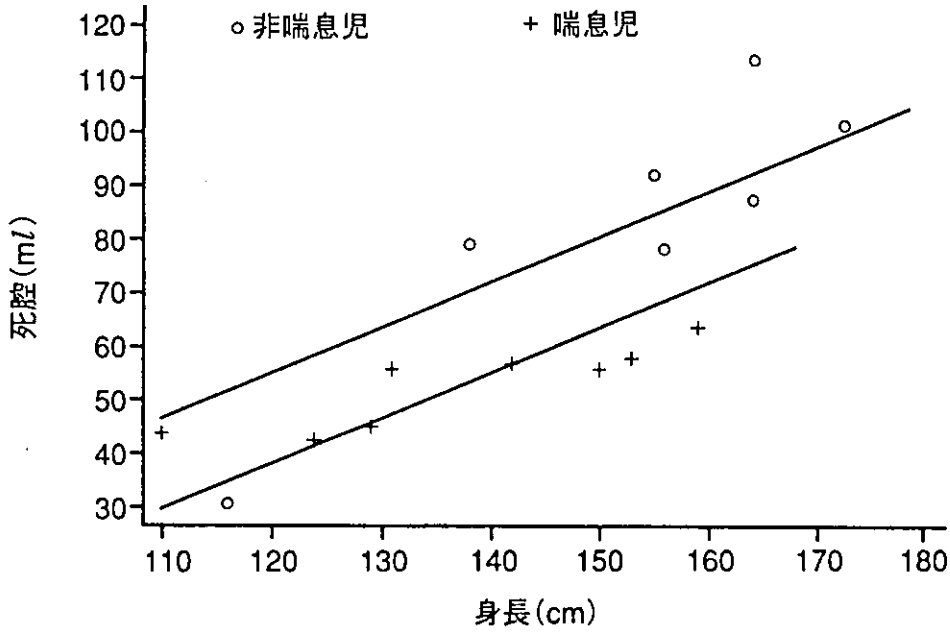

図 2.2 喘息児と非喘息児の傾きが同じ

2. 勾配は同じだが切片が異なる場合

モデルは，次のようになる．

$$\text{死腔} = \beta_0 + \beta_{\text{身長}} \times \text{身長} + \beta_{\text{喘息}} \times \text{喘息} \qquad (2.3)$$

図 2.2 に示した．モデル 2.3 からわかるように，係数 $\beta_{\text{喘息}}$ は勾配が $\beta_{\text{身長}}$ である 2 本の平行直線に関する切片の差を表す．身長に左右されない，喘息児と非喘息児とのあいだの死腔に関する差である．別の言い方をすれば，身長を**考慮** (allowing for) したうえでの差である．したがって，標本の中で喘息児と非喘息児の死腔の違いが身長の違いによると考えたならば，これは採用できるモデルの一つである．この種のモデルは**共分散分析** (analysis of covariance) と呼んでいる．医学論文にはよく出てくる．勾配が 2 グループで共通ということが大切な仮定である．

式 (2.2) と (2.3) で同じ記号を使っているが，それぞれのモデルに当てはめると $\beta_{\text{身長}}$ の推定値は異なる．

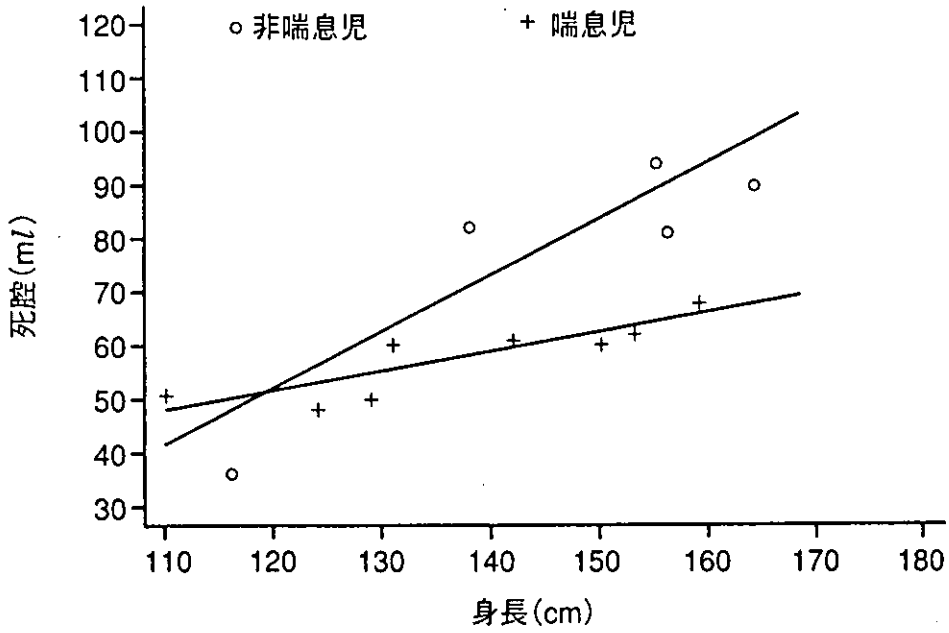

図 2.3　喘息児と非喘息児，それぞれの直線

3. 勾配と切片がそれぞれのグループで異なる場合

これをモデル化するためには，第3の変数 x_3＝身長×喘息が必要になる．x_3 は被験者が喘息であれば身長に等しくなり，喘息でなければ0を取る．変数 x_3 は喘息の状態と身長のあいだの**交互作用** (interaction) を表す．死腔と身長の関係を示す勾配が，喘息か喘息でないかでどのくらい変わるかを表す．

したがってモデルは，次のようになる．

$$\text{死腔}=\beta_0+\beta_{\text{身長}}\times\text{身長}+\beta_{\text{喘息}}\times\text{喘息}+\beta_3\times\text{身長}\times\text{喘息} \qquad (2.4)$$

図 2.3 で示したように，喘息児と非喘息児それぞれに勾配が異なる．2つの直線は以下のようになる．

非喘息児の場合には

$$\text{グループ}=0：\text{死腔}=\beta_0+\beta_{\text{身長}}\times\text{身長}$$

喘息児の場合には

$$\text{グループ}=1：\text{死腔}=(\beta_0+\beta_{\text{喘息}})+(\beta_{\text{身長}}+\beta_3)\times\text{身長}$$

となる．このモデルの$\beta_{身長}$の解釈は，モデル（2.3）とは異なり，非喘息児における直線の勾配の予測値である．一方，喘息児における直線の勾配は$\beta_{身長}+\beta_3$になる．β_3が与えられれば，喘息児と非喘息児の勾配の違いが分かる．

2.3.2 2つとも連続な独立変数の場合

表2.1のデータをもとに，両方の独立変数が連続である場合を考えてみよう．まず，身長と年齢の両方ともが死腔の予測に重要なのかどうかをみてみよう．

モデル式は次のようになる．

$$死腔=\beta_0+\beta_1\times身長+\beta_2\times年齢$$

このモデルの解釈は，前のものより意外に手ごわい．そして図式的に示すことも難しい．被験者の全員が同じ年齢だが，みんな違った身長であると想像してみてほしい．その時，被験者の年齢にかかわらず，身長が1 cm大きくなるごとに死腔がβ_1 m*l*だけ大きくなると予測できる．こんどは被験者が全員同じ身長だが，異なった年齢の集団を想像してみてほしい．その時，被験者の身長にかかわらず，年齢が1歳増えると死腔がβ_2 m*l*大きくなると予測できる．このモデルのすぐれた点は，ちょうど同じ年齢または同じ身長の被験者がいなくても，これらの係数を適切に推定できることである．

このモデルは，2.2節で述べた「予測」のところで通常使用される．

2.3.3 2つともカテゴリー独立変数の場合

表2.1の喘息状態をコード化するには，**ダミー変数**（dummy）や**標識変数**（indicator variable）を用いる．喘息と非喘息という2水準だと，たった一つのダミー変数でよい．ダミー変数の係数は喘息と非喘息でのy変数の違いを表す．喘息を1で正常を0，または逆にコード化しても問題はない．係数の符号が変わるだけである．P値は同じである．とこ

表 2.2　3カテゴリーの変数をコード化する例

状態	x_1	x_2	x_3
喘息	1	0	0
気管支炎	0	1	0
正常	0	0	1

ろで，表には3つのカテゴリーがある．つまり喘息，気管支炎，そしてどちらでもない（つまり正常）の3つである．これらのカテゴリーは重複していない（すなわち喘息と気管支炎両方もつ子は一人もいない）．**表 2.2** は，3つの被験者グループがあるときのダミー変数のつけ方を示している．

ここでは3つの可能な対比がある．つまり「喘息」対「気管支炎」，「喘息」対「正常」，「気管支炎」対「正常」である．しかし，それらは全部が独立というわけではない．対比する2つが分かれば，3番目は分かってしまう〔もしも喘息や気管支炎でないなら，正常に**違いない** (must)〕．したがって，3つの対比から2つ，つまり回帰の中に投入する2つのダミー変数を選ぶ必要がある．もし3つの変数すべてを含めると，ほとんどの回帰プログラムは，x_1，x_2，x_3 の係数は**重複しています** (aliased)（すなわち相互に依存している）と親切に教えてくれ，自動的に変数の1つを省くだろう．回帰から省かれるダミー変数は，他のダミー変数の対照となるものであり，**ベースライン** (baseline) 変数と呼んでいる．もし表2.2で x_1 と x_2 を含む回帰から x_3 が省かれるとすると，その時 x_1 の係数は喘息と正常との間の死腔の違いになる．これは見方を変えれば，ベースラインの係数はゼロになっているということになる．

2.4　コンピュータ出力の解釈

線形回帰のコンピュータ出力の解釈について述べる．ほとんどの統計パッケージは，これと似た出力をしてくる．**「最小二乗原理」**(principle of least squares) を用いてモデルを当てはめる．これについては［付録 2］で説明しているが，誤差分布が正規のときは最尤法に等しい．

2.4.1　1つが連続，1つが二値の独立変数の場合

最初に新しい変数を作り，喘息に対して＜Asthma＝1＞，非喘息に対して＜Asthma＝0＞とする．喘息と身長の交互作用を表す新しい変数 AsthmaHt=Asthma×Height を作る．いくつかの汎用プログラムは，喘息を『因子 (factor)』あるいは『カテゴリー (categorical)』と設定し，モデルの中で Asthma*Height と指定すると，これらを自動的に行ってくれる．

表 2.3 にコンピュータプログラムを使用して，これらの変数を当てはめた結果を示した．

死腔に対して 3 つの独立変数，Height，Asthma，AsthmaHt を当てはめた．これはモデル (2.4) と同値であり，図 2.3 に示されている．コン

表 2.3　表 2.1 の身長と喘息状態およびこの二つの相互作用を死腔に当てはめた場合のコンピュータによる計算結果

Source	SS	df	MS	Number of obs＝15	
				F(3, 11)＝37.08	
Model	7124.3865	3	2374.7955	Prob > F＝0.0000	
Residual	704.546834	11	64.0497122	R-squared＝0.9100	
				Adj R-squared＝0.8855	
Total	7828.93333	14	559.209524	Root MSE＝8.0031	
Deadspace	**Coef.**	**Std. Err.**	**t**	**P>| t |**	**[95% Conf. Interval]**
Height	1.192565	.1635673	7.291	0.000	.8325555　1.552574
Asthma	95.47263	35.61056	2.681	0.021	17.09433　173.8509
AsthmaHt	−.7782494	.2447751	−3.179	0.009	−1.316996　−.239503
_cons	−99.46241	25.20795	−3.946	0.002	−154.9447　−43.98009