ピュータプログラムの出力は2つに分かれている．最初の部分は全体のモデル適合性についてである．F(3, 11)＝37.08は（統計学者フィッシャーにちなんで）F統計量として知られており，これは2つの**自由度**（degrees of freedom）に依存している．最初の自由度kは，（定数項β_0を除いた）モデル中のパラメータ数であり，この場合では3である．2番目の自由度は$n-k-1$から得られる数字であり，nは被験者数を表す．この場合は，15−3−1＝11となる．「定数項だけで説明変数なし」というのが真のモデルだとして，この3変数モデルで説明される変動が偶然ゆえに起こる確率がProb＞Fである(訳注1)．別のことばで言えば，モデル全体での有意性ということになる．この場合0.0000と表示されており，P＜0.0001と解釈する．つまり，3つすべての変数を**同時に**（simultaneously）当てはめたモデルは非常に有意ということである(訳注2)．しかし，これは個々の変数について言っているのでは**ない**（not）．重要な統計量はR^2値(訳注3)であり，これはモデルによって説明される元データの分散の割合である．このモデルでは0.91である．たった1つしか独立変数をもたないモデルでは，Swinscow[1]で述べたように，この数字は単純に相関係数の二乗となる．しかし，データ数と同じだけのパラメータをあてはめれば，意図的に適合度を良くすることができる．このことを考慮して，**自由度調節済み**（adjusted for degrees of freedom）R^2を計算する．これは$R^2_a=1-(1-R^2)(n-1)/(n-k)$で計算でき，この場合では0.89である．root MSEは，『残差の平均平方誤差（residual mean square error)』を意味する．この場合，その値は8.0031である．これは方程式（2.1）のσの推定値である．そして，左の表のresidual MS（残差の平均平方）の平方根として計算できる．つまり$\sqrt{64.0497}=8.0031$となる．

訳注1：3変数モデルで説明される変動が，定数項モデルで説明できる可能性がP値である．

2：定数項モデルに比べて有意に優れるということ．

3：表の中のR-squaredのこと．

出力の第2の部分では，モデル中のそれぞれの係数を検討している．身長と喘息状態の交互作用項が有意（P＝0.009）であることが分かる．勾配の**違い**（difference）は－0.778単位（95% 信頼区間：－1.317～－0.240）である．モデルから落とすべき項目は一つもない．もし主項目の一つ，喘息または身長が有意でないとしても，もし交互作用の項が有意なら，モデルから除くことができない．なぜなら，交互作用は主効果――この場合には喘息と身長――がなければ解釈できないからである．

最も合う二つの直線は次のようになる．

非喘息の場合：死腔＝－99.46＋1.193×身長

喘息の場合　：死腔＝(－99.46＋95.47)＋(1.193－0.778)×身長

＝－3.99＋0.425×身長

このように死腔は，喘息児では非喘息児より身長とともによりゆっくりと大きくなるようである．

明白なことは，これらの方程式の切片は無意味だということである．なぜなら，それは被験者の身長が0と仮定した場合の死腔の予測値なので，全く意味がないからである．

2.4.2 2つの独立変数がともに連続の場合

ここでは身長と年齢，あるいは両方が死腔の予測に重要かどうかを考えたい．分析結果は**表2.4**のとおりである．

式は次のとおりとなる．

死腔＝－59.05＋0.707×身長＋3.045×年齢

このモデルの解釈は2.3.2で説明した．この結果の特徴に注目してほしい．モデル全体では有意（P＝0.0003）であるにもかかわらず，身長と年齢の係数は両方とも有意でない（それぞれP＝0.063，P＝0.291）．これは，年齢と身長が強く関係しているために起きる．またモデルの全体の当てはまりを見ることの重要性を物語っている．一方を除けば他方がモデルの重要な予測因子として残るだろう．年齢を除いても，調整済

表 2.4　表 2.1 の年齢と身長を死腔にて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(2,12)	=	17.29
Model	5812.17397	2	2906.08698	Prob > F	=	0.0003
Residual	2016.75936	12	168.06328	R-squared	=	0.7424
				Adj R-squared	=	0.6995
Total	7828.93333	14	559.209524	Root MSE	=	12.964

Deadspace	Coef.	Std. Err.	t	P>| t |	[95% Conf. Interval]	
Height	.7070318	.3455362	2.046	0.063	−.0458268	1.45989
Age	3.044691	2.758517	1.104	0.291	−2.965602	9.054984
_cons	−59.05205	33.63162	−1.756	0.105	−132.329	14.22495

Bootstrap statistics

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]		
Height	1000	.7070318	−.0080937	.3313434	.056823	1.357241	(N)
					.0793041	1.312535	(P)
					.0845788	1.31849	(BC)
Age	1000	3.044691	.3040586	3.399811	−3.6269	9.716281	(N)
					−2.586633	10.66853	(P)
					−2.986388	10.29889	(BC)

N = normal, P = percentile, BC = bias-corrected

み R^2 は，あまり影響されないことに注目してほしい（年齢と身長がある場合 $R^2=0.6995$，身長だけの場合 $R^2=0.6944$ を比較してほしい）．このことは，身長がより良い予測因子であることを示唆している．

2.4.3　ブートストラップ推定の使用

表 2.4 の下半分には，**ブートストラップ** (bootstrap) という，コンピュータを駆使した方法で得られた結果を示した．この方法を用いると，より安定した回帰係数の標準誤差が推定できる．ブートストラップの基礎については［付録 3］に記載した．

この方法は，この本で述べているような方法よりも分布に依存しない．多数回にわたりデータをサンプリングし，その都度回帰式を計算することになる．例えば，残差をプロットしてみてその分布がひどく非対称性

であったら，この方法を使う．いつもとは異なる3種類の推定値をコンピュータは計算してくる．それらは正規推定値（N），パーセント点推定値（P），バイアス補正推定値（BC）である．最後の推定値を勧める．ブートストラップ法で求めた身長推定値の標準誤差推定値は，従来の推定値よりわずかに小さいことが分かるだろう．したがって，信頼区間は0を含まなくなる．ブートストラップ法による年齢の標準誤差はもっと大きくなる．これにより，身長はより強い予測因子であるという先の結論が確かめられた．

2.4.4 カテゴリーの独立変数の場合

3グループの死腔の平均値（ml）が，正常97.33，喘息52.88，気管支炎72.25の場合で考えてみよう．分析結果は**表2.5**に示されている．ここで，2つの独立変数は表2.3のx_1，x_2である．前に述べたようにチェックすべき重要な点は，個々の対比を見る前に，モデル全体が有意

表2.5 表2.1の二つのカテゴリー変数を死腔にて当てはめた場合のコンピュータによる計算結果

Asthma and bronchitis as independent variables

Number of obs=15, F(2,12)=7.97, Prob>F=0.0063
R-squared=0.5705 Adj R-squared=0.4990

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Asthma	−44.45833	11.33229	−3.923	0.002	−69.14928	−19.76739
Bronch	−25.08333	12.78455	−1.962	0.073	−52.93848	2.771809
_cons	97.33333	9.664212	10.072	0.000	76.27683	118.3898

Asthma and Normal as independent variables

Number of obs=15, F(2,12)=7.97, Prob>F=0.0063
R-squared=0.5705, Adj R-squared=0.4990

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Asthma	−19.375	10.25044	−1.890	0.083	−41.7088	2.9588
Normal	25.08333	12.78455	1.962	0.073	−2.771809	52.93848
_cons	72.25	8.369453	8.633	0.000	54.01453	90.48547

かどうかを見ておくということである．ここではProb＞F＝0.0063なので，モデル全体として高度に有意である．それぞれの対比を見ると喘息の係数が－44.46であり，これは正常と喘息の平均の違いを意味する．この標準誤差は11.33で高度に有意である．気管支炎の係数は－25.08で，これは気管支炎と正常との違いであるが，これについては有意でない．つまり気管支炎と正常との平均の死腔には，有意な差がないことを示している．

喘息と気管支炎を比較したいとすると，どちらかをベースラインにする必要がある．ここでは気管支炎をベースラインと考え，独立変数をx_1とx_3とすると，出力結果は表2.5のようになる．予測されたように，Prob＞FやR^2の値は先程のモデルと同じである．なぜなら，これらの数値はモデル全体についてのものであり，先程のモデルとの違いはパラメータの取り方だけだからである．しかし気管支炎との対比で言えば，喘息と気管支炎との係数は－19.38で標準誤差は10.25である．しかし，これは有意ではない．

こうして見てくると唯一の有意な差は，喘息と正常の間にあることが分かる．

この方法は，**一元配置分散分析**（one-way analysis of variance）と呼ばれている．Swinscow[1]の中で述べたように，これはt検定を一般化したものである．この方法とただ単にt検定を2回くり返す方法，例えば――喘息と正常，気管支炎と正常――を行う場合と何が違うのかと疑問に思うかもしれない．しかし実際には，分散分析は2つの特別な改良をしている．第1の改良点は，Swinscow[1]の中で触れた多重検定の問題を考えて，全体のP値を計算している点である．基準値に対して検定をくり返すと，第1種の過誤の機会が増す．F検定のP値はこのことを考慮に入れている．もし全体で有意ならば，対比のいくつかが有意ということになる．第二の改良点は何かというと，t検定を計算するとき併合した標準誤差を使っているということである．t検定では標準誤差を2グ

ループから導くが，分散分析ではすべてのグループから計算する．こうすることで多くの被験者に基づくことになり，より正確なものとなる．

2.5 重回帰の実際

2.5.1 共分散分析

論文中でモデル (2.3) がよく見られることは，すでに述べたとおりである．臨床試験への応用例として Llewellyn-Jones ら[3] の結果について考えてみよう．結果の一部は**表 2.6** のとおりである．この研究は，65 歳以上の 220 人のうつ病患者を対象に，集団介護が有効であるかを調べるランダム化比較対照試験であった．うつ病は Geriatric Depression Scale (高齢者うつスケール) を用い，ベースライン (試験開始時) と盲検下で追跡し 9.5 ヵ月後に評価した．図 2.2 は解釈を助けとなるだろう．ここでは y は 9.5 ヵ月治療後のうつスケール (連続数)，x_1 はベースライン値，そして x_2 は，1 を介入，0 を対照とするグループ変数である．

標準化回帰係数 (standardised regression coefficient) の定義で一般的なものはないが，この場合は x をその標準偏差で割った値とする．したがって，標準化回帰係数を解釈すると，x が 1 標準偏差増加したときの y の変化量になる．ベースラインの値は，追跡後の値と強い関連をもつことが分かる．ベースラインがどういう値であっても，介入により，平均で −1.87 単位 (95% 信頼区間：0.76〜2.97) 低い値を示していた．

表 2.6 追跡時高齢者うつスケールに影響する因子

変数	回帰係数 (95% 信頼区間)	標準化回帰係数	P 値
ベースライン値	0.73(0.56〜0.91)	0.56	<0.0001
治療群	−1.87(−2.97〜−0.76)	−0.22	0.0011

この分析では，治療効果はすべての被験者で等しく，ベースラインの値とは関係しないと仮定している．この仮定は，先に述べた方法で調べることが可能である．二つの集団のベースライン値が似ていれば，ベースラインの値を含めなくても治療の比較へは影響ないだろうと思われる．しかし，多くの場合ベースラインの値を含めたほうがよい．なぜなら，治療効果の推定程度を増すことになるからである．すなわち，ベースラインを共変量として含めたほうが，治療効果の標準誤差は小さくなるだろう．

2.5.2 2つの連続変数が独立変数の場合

Sorensen[1]らは，BMIを測定した18～26歳の男性4,300人を前向きに研究した．彼らは，成人のBMIと出生時体重・身長が関係しているかを調べた．可能性のある交絡因子には妊娠週数，第何子か，母親の婚姻状態，年齢，そして職業を考えた．(交絡を考慮した)線形重回帰を行ったところ，出生時体重（250 g単位でコード化）とBMIのあいだに回帰係数0.82でSE 0.17と有意の関係が認められた．しかし出生時身長(cm)とBMIのあいだには，回帰係数1.51でSE 3.87のため有意な関係は認められなかった．このことは出生時体重が250 g増加するごとに，BMIは平均0.82 kg/m^{2}増加するということである．妊娠期間などの因子を考慮しても，出生時体重に影響する『子宮内の（in utero)』因子が，成人になっても依然としてBMIへ影響していると，著者らは考えている．

2.6 モデルの基礎となる仮定

モデル選択に関していくつかの暗黙の仮定がある．最も基本的な仮定は，モデルが**線形**（linear）だということである．これは，変数xの初期値にかかわらず，変数xが1単位増加すると変数yが一定量増加す

るという仮定である.

x が連続の場合，これを調べるためのいくつかの方法がある.

- 連続な独立変数が 1 つのとき，最も簡単なのは，x に対する y の散布図を描くという視覚的な方法である.
- 変数 x を変換する〔$\log(x)$, x^2, $1/x$ が一般的である〕. 二種類の変換を比較するための簡単な有意性検定はないが，R^2 値が良い指標となる.
- モデルに線形項（x）と二乗の項（x^2）を入れてみる. このモデルでは，2 つの連続変数，x と x^2 を当てはめたことになる. x^2 項が有意であれば，直線性の欠如を示すことになる.
- 例えば 5 分位法などのように，x をいくつかのグループに分ける. 5 分位の中の大きいほうの 4 グループにそれぞれダミー変数を当てはめ，それらの係数を調べる. 線形関係があれば，それらの係数は直線的に増加する.

もう一つの基本的な仮定は，誤差項はそれぞれ独立だということである. これが成り立たなそうな例として時系列データがある. 逐次データの誤差が独立かどうかを簡単にチェックするには，残差が相関しているかどうかを調べる **Durbin-Watson 検定**（Durbin-Watson test）がある. これは多くの統計パッケージで利用できる. さらに詳細は，時系列分析を扱っている第 6 章を見てほしい. 独立性が欠如するもう一つの例は，測定の主要な単位が個人の場合である. つまり個人に何回も測定されているにもかかわらず，あたかも別人から測定されたように扱う場合である. これは**反復測定**（repeated measures）の問題である. 似たようなことは，患者集団へ割り付けたときに起こる. これについては，反復測定を扱った第 5 章を参照してほしい.

モデルというのは，誤差項と変数 x とは独立だと仮定している. さらに，誤差項の分散は一定であると仮定している〔そうでないと**不等分散**（heteroscedascity）になる〕. よくあるのは，変数 x が 1 単位増加す

るにつれて誤差が増加する場合である．仮定をチェックする一つの方法は，残差 e_i を y 軸にし，各独立変数を x 軸にとりプロットするものである．x 軸としてモデルによる当てはめ値をとりプロットするのもよい．もしモデルが正しければ x 軸に関して一様に残差がプロットされ，いかなるパターンも示さないはずである．もし逸脱するとしたら，一般に残差が扇形に広がる場合である．つまり，変数 x が大きくなるにつれてばらつきが大きくなる場合である．この理由としては非直線性が考えられ，変数 x を変換すれば問題は解決するかもしれない．

最後の仮定は誤差項が正規分布することである．これをチェックするには，残差のヒストグラムを描けばよい．ただこの方法を使うと，真の残差 ε_i よりも実際の残差 e_i のほうが正規分布しがちになることに注意したい．正規性の仮定が重要なのは主に，係数の信頼区間推定に正規理論を使用するためである．しかし運良く適当な数の標本数があれば，正規性から逸脱していても安定した推定が得られる．だから，正規性から多少逸脱してもよい．また，[付録 3] に記載したブートストラップ法を使うこともできる．

解析する主目的は，「仮定を検定することでは**「なく (not)」**，関連性を評価すること」だと覚えていてほしい．**「仮定が完全に満たされないときでも」** (even when the assumptions are not perfectly satisfied)，多くの場合有益な結論が得られるものである．

2.7 モデルの感度

モデルの感度とは，データのサブグループによってどのくらい推定値が影響されるかということである．データの一部分か一つだけデータを削除してモデル（モデル 2.2）を当てはめ，推定値 b_0 と b_1 が大きく変わったとしよう．これは大変なことである．なぜなら広く適用できるモデルを探しており，患者のサブグループ別に異なったモデルを見つけた

いわけではないからである.

2.7.1 残差，てこ比，影響

個々の測定値のモデルへの感度を調べるには三つの重要なことがある．それらは**残差** (residuals)，**てこ比** (leverage)，**影響度** (influence) である．残差とは，測定データとモデル当てはめ値との差，つまり $e_i = y_i^{obs} - y_i^{fit}$ である．大きな残差を示すデータのことを，**外れ値** (outlier) と呼ぶ．一般に外れ値は推定値に影響を及ぼすかもしれないので，注意を払う必要がある．しかし，影響を及ばさない外れ値もある．

外れ値を調べる別の方法としては，x_i の値が x の集団から離れているか見ることがある．1 変数でいえば，x_i が $\bar{x}$ (平均) から離れているかを見ればよい．x に対して y の散布図を描いたとき，多くのデータが左下隅にあり，1 つだけが右上にあったとしよう．このデータは変数 x と y の関係に特異な性質をもつかもしれない．データに当てはめた回帰直線が，孤立した点に近づくこともあろう．ときにはそこを通るかもしれない．そんなときには，この孤立した点は大きな残差を示さない．しかし，この点を削除すると，回帰係数はひどく変わるかもしれない．そのようなデータは，**高いてこ比** (high leverage) だという．それは数量化され，通常 h_i で表す．h_i の値が大きいほど，てこ比も大きい．

影響データとは，推定に大きな影響をもつデータのことである．これを調べるには，そのようなデータを含む場合と含まない場合でモデルを当てはめ，回帰係数への影響をみる．b_0 や b_1，あるいは SE (b_1) などへ，大きな影響を及ぼすデータが見つかるだろう．通常，そのようなデータを含むときと含まないときでの係数の差が出力される．その際，推定標準誤差で標準化する．問題は，パラメータごとに影響力データが異なるかもしれないことである．ほとんどの統計ソフトではルーチンに残差，てこ比，影響度を出力する．解析する人はそれを見て，重大な症例を見抜かないといけない．しかし，ただ影響しているとか残差が大きいとい

うだけでは，削除すべきではない．そのようなデータについては，測定ミスや転記ミスを注意して調べておきたい．データを適切に分析すれば，個々のデータの感度がわかるはずである．

2.7.2 コンピュータ分析：モデルの点検と感度

表 2.1 の死腔，年齢，身長のデータを用いて，モデルの点検と感度について説明する．

図 2.1 から，死腔と身長との関係が線形だと確かめることができる．死腔と年齢についても同じようなグラフを描ける．標準的な診断プロットとは，モデルから得られる値との残差プロットのことである．表 2.3 であてはめたモデルについては，**図 2.4** に示したとおりである．明らかなパターンはみられないので，誤差項は相対的に一定であり，モデルの直線性も確認できた．

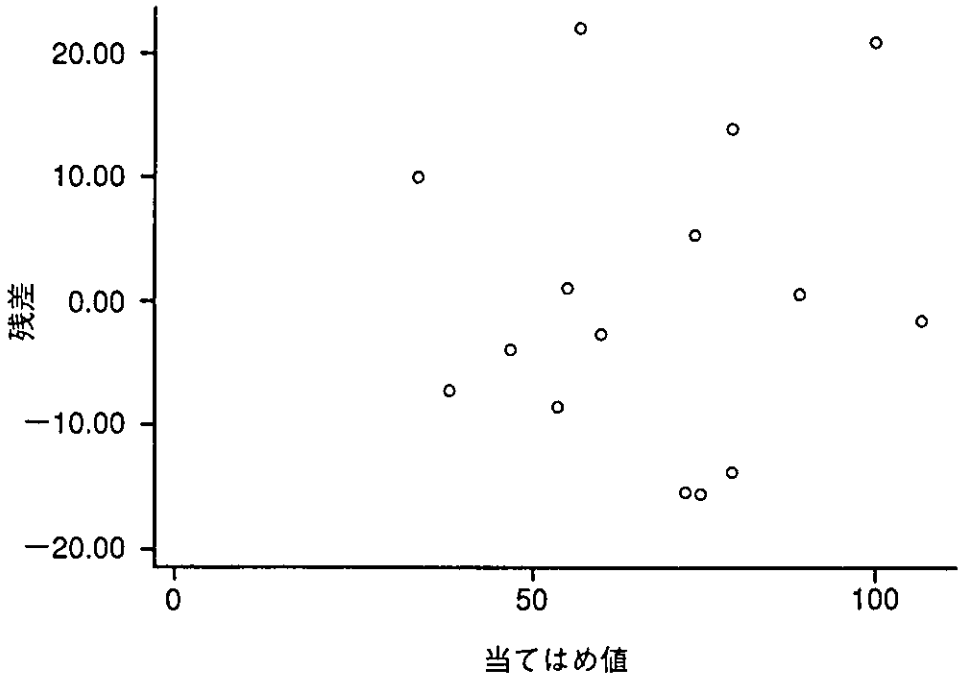

図 2.4 表 2.4 の年齢と身長を独立変数とした回帰モデルでのモデルから得られた値に対する残差のグラフ

表 2.7 表 2.4 でモデルを当てはめた場合の診断

	身長	年齢	残差	てこ比	影響度(年齢)	影響度(身長)
1.	110	5	10.06	0.33	0.22	−0.48
2.	116	5	−7.19	0.23	−0.04	0.18
3.	124	6	−3.89	0.15	−0.03	0.08
4.	129	7	−8.47	0.15	−0.14	0.20
5.	131	7	1.12	0.12	0.01	−0.02
6.	138	6	22.21	0.13	−0.52	0.34
7.	142	6	−2.61	0.17	0.08	−0.06
8.	150	8	−15.36	0.08	0.11	−0.14
9.	153	8	−15.48	0.10	0.20	−0.26
10.	155	9	14.06	0.09	0.02	0.07
11.	156	7	5.44	0.28	−0.24	0.25
12.	159	8	−13.72	0.19	0.38	−0.46
13.	164	10	0.65	0.14	0.00	0.01
14.	168	11	18.78	0.19	0.29	0.08
15.	174	14	−5.60	0.65	−0.79	0.42

表 2.7 には診断統計量を示した．**影響度** (influence) の統計量として，年齢と関係するものを **inf_age**，身長と関係するものを **inf_ht** で表した．予想されたとおり，最年少の（最も背が低い）子どもと最年長の（最も背が高い）子どもで，最大てこ比を示した．最大の残差だと，てこ比は小さくなることに注意しよう．なぜなら，大きなてこ比を示すデータは，直線を近づける傾向があるためである．

年齢に最も影響されるのは最年長の子どもであり，それらを除去すると標準化回帰係数が 0.79 単位も変わる．身長に最も影響する子どもは，背が一番低い子どもである．しかし，十分な理由なくしてどちらの子どもも除去すべきではない（十分な理由とは，子供が何か関連する病気にかかっていると分かったときである．例えば，嚢胞性線維症など）．

2.8 ステップワイズ回帰

独立変数がたくさんあるとき，変数 y を予測する変数の組み合わせでどれが一番良いかと思うだろう．こういったとき**ステップワイズ回帰**

(stepwise regression) を使えばよい．これはいくつかの統計パッケージで利用できる．**ステップダウン** (Step-down) あるいは**バックワード回帰** (backwards regression) とは，最初にすべての変数を含め，その後で有意でない変数を除外していく方法である．**ステップアップ** (Step-up) あるいは**フォワード回帰** (forwards regression) とは，最初に全体平均だけで，重要度に応じてモデルに変数を加えていく方法である．**ステップワイズ** (Stepwise regression) 回帰はこの二つをミックスした方法であり，モデルに入れる際のP値と捨てる際のP値を指定できる．通常はモデルから除く場合（例えば0.05）より，入れる場合のほうが大きなP値（例えば0.1）を設定する．というのは，変数一つ一つでは予測性が高くなくても，いくつかいっしょに見ると予測性が上がるからである．このためステップダウン回帰が好まれている．結果変数として，足をひきずる程度をとった例を考えてみよう．左足も右足も予測性はないが，両足の長さの違いは予測性がある．ステップワイズ回帰は**探索的分析** (exploratory phase)（第1章参照）でよく使う．多くのデータからいくつか予測因子を見つけたり，さらにデータを収集してその関連性を検証したりするときに使う．

ステップワイズ回帰を使うときに，いくつかの問題がある．

- 何度も検定したことを考慮していないので，P値は意味がない．ステップアップやステップダウンなどの別の手法では，多分違ったモデルになるだろう．経験上，2度目の分析をした場合，1度目と同じモデルになることはめったにない．これを解決する一つの方法は，大きなデータセットを2つに分け，両方別々にステップワイズ法を行うものである．両方のデータセットに共通する変数を見つけ，それらを含めたモデルで全体のデータをもう一度解析する．
- 大きなデータセットには欠損値が含まれているものである．ステップワイズ回帰では，通常どの (any) 変数についても欠損値のない被験者データしか使われない．そうすると最終モデルに含まれた変数はわ

ずかかもしれない．モデルの当てはめを再度行うと，パラメータが変わってしまうこともある．なぜなら今当てはめようとしているモデルは，ほんの少しの変数についてだけ欠損値がない被験者に当てはめようとしているからである．このため最終モデルを当てはめたデータは，オリジナル[訳注]のものよりかなり大きいかもしれない．

- カテゴリー変数がダミー変数でコード化されている場合，そのいくつかは当てはめの過程で除かれてしまうかもしれない．このため他の変数の解釈まで変わってしまう．例えば，表2.2のx_1とx_2でモデルに当てはめたが，x_2を除くとしよう．このときx_1の解釈としては，気管支炎の喘息児と健康な子どもとの違いを無視した結果になる．

このようにステップワイズ回帰は探索的分析では有用だが，**検証的分析**(confirmatory) には向かない．

2.9 重回帰の結果を報告するには

- モデルに関する調整済みR^2と，主な独立変数については回帰係数および標準誤差または信頼区間は記載しなさい．
- 信頼区間をブートストラップ法で推定したら，使用した方法(例えば，バイアス補正) と反復回数を記載しなさい．
- 主な従属変数が一つだけなら，最もよく当てはまった直線と実際のデータ点をプロットしなさい．
- モデルの基礎となる仮定を，どのように検証したかを述べなさい．特に線形性の適合度は忘れないように．
- 実施したすべての感度分析を記載しなさい．
- モデルに含んだ**すべての** (all) 変数を記載しなさい．ステップワイズ回帰では，モデルに投入した**すべての** (all) 変数を記載しなさい．
- もし交互作用項をモデルに含むなら，主効果が含まれ**ているか** (must)

訳注：すべての変数が欠損値でない被験者に限ったデータ．

を確認しなさい.

2.10 重回帰分析結果の読み方

1.11 節の中の要点に加えて

- R^2 値をよく見なさい. 例数が大きいとモデルの係数が有意になりがちだが, 結果変数のばらつきのわずかしか説明していない場合がある. そうであれば, あまり結果を予測するには役立たないだろう.
- モデルは線形でよいか. 勾配が平坦になるような閾値はないか.
- 外れ値や影響データを見つけて, それらの対処法は述べたか. それらはどのように扱われているか.
- 共分散分析では, それぞれのグループで勾配は等しいと**仮定している** (assumes). これは満たされると考えられるか, また検証されたか.

[よく質問されること (FAQ)]

1. ダミー変数はどのようにコード化したらよいのか?

二値変数が一つしかないときは, 0 と 1 とコード化したダミー変数を用いる. アンケートでは通常 1 と 2 にコード化するが, これでも係数の推定値や P 値は変わらないが, 切片の値が変わる. なぜなら今 1 を割り当てたグループの値は $a+b$ で, 2 を割り当てたグループの値は $a+2b$ になるからである. 例えば, 図 2.2 で「喘息」の有無を 0 と 1 でコード化した場合, 喘息の回帰係数は −16.8 で切片は −46.3 となる. もし 1 と 2 でコード化すると, 喘息の回帰係数は −16.8 と変わらないが, 切片は (−46.3−16.8) = −63.1 となる. ダミー変数を −1 と +1 とコード化すると (例えば SAS ではこうするが), P 値は変わらないが, 係数は半分になる.

3 水準のカテゴリー変数では, 2 つのダミー変数でコード化する. 先に見たように, 全体の F 統計量はダミー変数をどう設定しても

変わらない．しかし，水準1か水準3のどちらを省略するかによって，水準2の係数は変わる．

2. 順序の独立変数はどのように扱ったらよいか？

ほとんどの統計パッケージでは，回帰モデルの予測因子（つまり変数X）は，連続か二値だと仮定している．このときいくつかの選択肢が考えられる．

（i）予測因子をあたかも連続データのように扱う．これは，カテゴリーに順序があると考えてモデルに組み入れることを意味する．しかし，X の変化量に対する y の変化量は一定であるという仮定が必要となる．

（ii）予測因子をカテゴリーとして扱う．つまり，一つのカテゴリーをベースにして他のすべての変数にダミー変数を当てる．こうすると予測因子の順序はなくなるが，線形性に関する仮定は必要なくなる．

（iii）変数 X を二値に分けて，再コード化する．例えば，X がある特定の水準かそれ以上なら1，そうでなければ0とする[訳注]．分岐点は，そのデータに最良のあてはまりを示すからではなく，別のデータから得られた理由により選ぶべきである．

この3つのうちどれを選ぶかは，要因の数による．データ数が多いと，選択肢（ii）で順序を無視したことによる情報の損失はそれほどでもない．変数 X が交絡因子や重要な因子でなければなおさらである．例えば，X が10年間隔で区分された年齢群なら，変数 y との線形関係を仮定するよりも，ダミー変数を当てはめるほうがよいかもしれない．

3. 重回帰を使うとき仮定で何が問題となるか？

重回帰の基礎となる仮定は，多くの場合検討されていない．それは，ひとつには調査者がそれは満たされていると確信しているから

訳注：本文には y とあったが，誤りと判断し X と修正した．

であり，また軽度の逸脱は分析結果に影響することが少ないからである．しかし，独立性の逸脱が経験上（反復測定や時系列から得られたデータ）明らかな場合には，最初からこれを調整する必要がある．線形性は推定に重要であり，独立変数を変換して当てはめることで調べられるだろう．分散が一様でなかったり，正規性が欠如すると標準誤差に影響し，従属変数を変換する必要が出てくる．正規性からの逸脱が最もよくわかるのは，外れ値が見つかった場合である．その場合は慎重に調べなければならない．特にそのデータが大きなてこ比を呈すような場合は，なおさらである．

4. 交絡因子は検定をすると有意でないが，これを分析に含めるべきか？

交絡因子と考えてもおかしくない，いくつかの変数（例えば年齢や性別）がある．このような因子はどんな解析を行っても有意になるかもしれない．これらの因子は分析に入れておくべきである．たとえ結果と有意の関係がなくても，主たる独立変数の効果に影響しているかもしれないからである．

5. もし従属変数が0か1だったらどうするか？

従属変数が0か1の場合，線形回帰から得られた係数は**線形判別関数**（linear discriminant function）になる．この場合，残差の正規性に関する仮定が破られても，グループを判別するのには役に立つ．しかし，こうした判別には**ロジスティック回帰**（logistic regression）（第3章）を使うのが普通である．

[多肢選択問題]

1. Rossら[5]は，アメリカの282の大都市地域とカナダの53の大都市地域から，労働者の男性死亡率と収入の中央値シェア（50%から入ってくる収入の割合）を回帰させた．それぞれの地域での収入の中央値シェアは説明変数とした．勾配の違いは有意で

あり（P<0.01），R^2=0.51であった．

モデルは，次のように書ける．

$$y_i = a + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + b_4X_{4i}$$

y_i は大都市地域 i (i=1...335) の100,000人当たりの死亡数

X_{1i} は，値1がアメリカ，0がカナダ

X_{2i} は，(上で定義された)地域 i の収入の中央値シェアの割合

$X_{3i} = X_{1i} \cdot X_{2i}$ (X_{1i} と X_{2i} の積)

X_{4i} は，地域 i の収入の中央値

次の文が正しいか誤りかを考えなさい．

(i) 死亡率は正規分布すると仮定している．

(ii) 勾配を比較する検定は自由度330のt検定である．

(iii) 死亡率と収入中央値との関係はアメリカとカナダで異なると仮定している．

(iv) 死亡率と収入中央値の割合との関係は線形だと仮定している．

(v) 残差変動は，アメリカとカナダで同じだと仮定している．

2. 重回帰式が $y = a + b_1X_1 + b_2X_2$ について，以下の文章が正しいか誤りかを考えなさい．

(i) 独立変数 X_1 と X_2 は連続でなければならない．

(ii) てこ比は，y の値に依存する．

(iii) 勾配 b_2 は X_1 の値に影響されない．

(iv) カテゴリー変数 X_2 が，3カテゴリーであれば，2つのダミー変数でモデル化できる．

(v) データが100個あれば．b_1 を検定するときの自由度は97である．

[演習問題の解答]

1. (i) 誤り（そうではなく，収入シェアの中央値，収入の中央値，国別で補正したときの，残差が正規という意味だから），(ii) 正解(330=282+53−5なので)，(iii) 誤り（シェアの中央値は異なると仮定

しているが，シェアの中央値との関係は同じだから），（iv）正解，（v）正解

2. （i）誤り（X_1 と X_2 は離散変数なので），（ii）誤り（それは X_1 と X_2 に依存するため），（iii）誤り（X_1 の値が変わると X_2 と y との関係も変わるので，b_2 へも影響するから），（iv）正解，（v）正解

■文献

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

2 Draper NR, Smith H. *Applied Regression Analysis, 3rd edn.* New York: John Wiley, 1998.

3 Llewellyn-Jones RH, Baikie KA, Smithers H, Cohen J, Snowdon J, Tennant CC. Multifaceted shared care intervention for late life depression in residential care: randomised controlled trial. *BMJ* 1999; **319**: 676–82.

4 Sorensen HT, Sabroe S, Rothman KJ, Gillman M, Fischer P, Sorensen TIA. Relation between weight and length at birth and body mass index in young adulthood: cohort study. *BMJ* 1997; **315:** 1137.

5 Ross NA, Wolfson MC, Dunn JR, Berthelot J-M, Kaplan GA, Lynch JW. Relation between income inequality and mortality in Canada and in the United States: cross-sectional assessment using census data and vital statistics. *BMJ* 2000; **320**: 898-902.

3 ロジスティック回帰

要　旨

二値の従属変数をモデル化するとき，適当な解析法とは**ロジスティック回帰**（logistic regression）というものである．Swinscow[1] の本では，二つの二値変数の関連性を検証する χ^2 検定を説明した．ロジスティック回帰とは，この χ^2 検定を一般化したものである．すなわち二値の従属変数に対して，一つあるいはそれ以上の独立変数を扱える．それらは二値であっても，カテゴリー変数（2 水準以上）や連続変数でもよい．ロジスティック回帰はまた，**ケースコントロール研究**（case–control studies）の解析にも有用である．マッチングを伴うケースコントロール研究では，**条件付きロジスティック回帰**（conditional logistic regression）という特殊な解析になる．

3.1　モデル

従属変数は，有りか無し（時には『成功（success）』か『失敗（failure）』ともいう）の**イベント**（event）で表される．したがって，調査においては疾病の存在がイベントになり，臨床試験では疾病の治癒がイベントになる．ここで，イベントに関連する因子を検討したい．イベントが起こるか否かを正しく予測することはできないので，実際に見ようとするのはイベント発生の**確率**（probability）に関連する因子である．

次の二つの状況を考えてみよう．

1．独立変数がすべてカテゴリーであり，そのため独立変数が同じ値をもつ人数を表にすることができる場合である．したがって，あるイ