存している．つまり，冠動脈死は年齢に強く依存する．このモデルには5つのパラメータが含まれているので，出力にも5つのパラメータがある．この出力結果は生存時間解析の結果と似ており，z 統計量は IRR とその標準誤差との比ではない．それは log(IRR)，つまり，〔log (1.43)〕の標準誤差に対する比である．

6.1.4 モデルの点検

モデルを点検する最も単純なやり方は，観測値とモデルにより予測された値を比較することである．予測値は，式 (6.2) に推定された係数を代入して求められる．従属変数は頻度なので，観測値と予測値を比較するために χ^2 検定が使えて，$X^2=12.13$，df＝4，P＝0.0164 を得る．この自由度が4というのは，予測値は5つの年齢グループと1つの喫煙グループに関する観測値に等しいという制約があるからで（他の喫煙グループの制約は先の年齢グループの制約からくる），6つの制約と10の観測値から自由度が4となる．このモデルがデータに適合していないという形跡がいくらかある．D の標準誤差は $\sqrt{E}$ なので，標準化残差は $(D-E)/\sqrt{E}$ と定義される．これらは表6.1に示され，大部分が−2と＋2の間に存在すると期待される．75～84歳年齢グループの非喫煙者を除いて，モデルに当てはまっていないとする系統的な乖離が明らかだという証拠はほんのわずかだけれども，その乖離は大きくはないと結論づけることができる．

さらに当てはめに利用できるのは喫煙×年齢の交互作用項があるだけだが，この項をモデルに含めるとこれは**飽和モデル** (saturated model，パラメータ数がデータ数と等しいモデルのこと，［付録2］参照）であり，その尤度比 χ^2 統計量は，上述の当てはまりからの乖離を表す χ^2 統計量に等しい．このようにして，喫煙は各年齢層で別々に冠動脈のリスクに影響を与えている形跡がある．

観測データがポアッソン分布で予測される以上に予測値から変移して

いれば，それは**超ポアッソン変動**（extra-Poisson variation）として知られている．これは第3章で述べた**超二項変動**（extra-Binomial variation）と同様である．したがって，コンピュータの出力結果中の標準誤差は妥当でないかもしれない．この現象は重要な共変量がモデルから省かれたためかもしれないし，別の一般的な説明としては頻度が相関していることに求められる．これは，別々の個体のグループ内における頻度よりもむしろ，年間の喘息発作回数のような1人の個体内の頻度を参照する時に生じる．このことは第5章で述べた変量効果モデルの話に繋がり，第5章で説明したように標準誤差が大きめに推定される傾向がある．現在では，ランダム効果ポアッソンモデルへのあてはめが可能な統計パッケージもある．λ_i 内の想定分布から外れる変動を許容する特別なモデルは，**負の二項回帰**（negative Binomial regression）モデルとして知られており，例えば，統計パッケージのSTATAが使える．

6.1.5 ポアッソン回帰の実例

Campbellら[4]は，イングランドとウエールズで1980～1995年の期間内の喘息による死亡を検討した．そこではその期間内の死亡に傾向があるのかどうかを検定するためにポアッソン回帰モデルを用い，1988年以来，15～44の年齢グループでは特に，年間約6%（95%信頼区間：5～7%）の低下傾向があるが，この低下傾向はより上の年齢グループでは認められないと結論づけた．

6.2 順序回帰

結果変数が**順序**（ordinal）である時には，はじめの方の章で述べた手法は適切ではない．一つの解決法はデータを二値化し，第3章のロジスティック回帰モデルを用いることである．しかし，この方法は有効ではないし，二値にするための分点をデータを見て選択する時に恐らくバイ

アスがかかる．順序回帰の主なモデルは**比例オッズモデル**（proportional odds model）または**累積ロジットモデル**（cumulative logit model）として知られている．これは，カテゴリーの確率よりもむしろ累積応答確率に基づく．

例えば，k 順序のカテゴリカル結果変数 y_j，$j=1,2,\ldots,k$，をもつ順位結果変数 Y を考えよう，そして共変量を $X_1,\ldots X_p$ とおこう．累積ロジットまたは比例オッズモデルは

$$\text{logit}(C_j)=\log\left[\frac{C_j}{1-C_j}\right]=\log\left[\frac{Pr(Y\le y_j)}{Pr(Y.y_j)}\right]=\alpha_j+\beta_1X_1+\ldots+\beta_pX_p, j=1,2,\ldots,k-1 \qquad (6.3)$$

または，等価的に

$$\Pr(Y\le y_j)=\frac{\exp(\alpha_j+\beta_1X_1+\ldots+\beta_pX_p)}{1+\exp(\alpha_j+\beta_1X_1+\ldots+\beta_pX_p)}\ j=1,2,\ldots,k-1 \qquad (6.4)$$

となる．ここで，$C_j=\Pr(Y\le y_j)$ はカテゴリー j での累積確率またはより小さい確率〔$_j=k$ に関して，$\Pr(Y\le y_j\mid X)=1$ であることに注意〕．ここで，記法の混乱を避けるために個体を指すための係数を用いないことにする．変数 α_j，$j=1,2,\ldots,k-1$ のセットにより，ロジスティック回帰モデル内で見られる切片項 β_0 を置換えたことに注意しよう．カテゴリーが $k=2$ である時にはこのモデルは方程式（3.1）と同一であり，ロジスティック回帰モデルとなる．2カテゴリー以上の時には，ベースのカテゴリーを除き，各カテゴリーについて別々の切片項を推定する．

回帰係数ベクトル β はカテゴリー i には依存しない．このことから，モデル（6.3）が共変量 x と Y の間の関係は i（応答カテゴリー）とは独立であると仮定していることがわかる．k 個のカテゴリーを通して同一の対数オッズ比をもつというこの仮定は，**比例オッズ仮説**（the proportional odds assumption）として知られている．

従属変数が連続変数ではあるが，その値は報告上グループ化されたと

思われる時は，比例オッズモデルは役に立つ．一方でその変数が取りうる値が制限された機器により，不完全に計測される場合もある．境界の間の区切りは時々**分点**（cutpoints）として知られている．応答 Y のコードを逆に設定（すなわち，y_1 を y_k として再コード化し，y_2 を y_{k-1} として再コード化し，等々）としても比例オッズモデルは不変である．次に，順位応答の隣接カテゴリーの折り畳み（例えば，y_1 を y_2 を組み合わせ，そして y_{k-1} を y_k を組み合わせる）の下でも比例オッズモデルは不変である．

ポアッソン回帰モデルの下で述べられた頻度データは順位変数とみなすことができる．しかし，頻度データは比尺度になるので，この場合順序回帰モデルは有効ではなさそうであるし，この事実は順序回帰モデルにおいて役に立たない（1.3 節参照）．

このモデルの解釈は，ロジスティック回帰モデルの解釈と全く同様である．連続変数と名義変数の共変量は独立変数に含めることができる．

6.2.1 コンピュータ出力の解釈

表 3.1 で与えられた授乳の長さは，1 ヵ月未満，1〜3 ヵ月，3 ヵ月以上として計測されたとする．このようにして分割点は 1 ヵ月目と 3 ヵ月目となる．このデータを**表 6.3** に示す．

結果変数は今や順位変数であり，これを反映した解析に用いることに対して感度が良い（sensible）．Swinscow[5] では，ノンパラメトリックなマン・ホイットニーの U 検定を用いた解析を示している．回帰モデル

表 6.3 子供に授乳した印刷工と農夫の妻の 1ヵ月未満，1〜3ヵ月，3ヵ月以上の数

妻	1 ヵ月未満	1〜3 ヵ月	3 ヵ月以上	計
印刷工の妻	20	16	14	50
農夫の妻	15	15	25	55
計	35	31	39	105

表 6.4 表 6.3 のデータに対する順序回帰分析の結果

Iteration 0： log likelihood＝ −114.89615
Iteration 1： log likelihood＝ −113.17681
Iteration 2： log likelihood＝ −113.17539

Ordered logit estimates　　Number of obs ＝ 105
　　LR chi 2(1) ＝ 3.44
　　Prob＞chi 2 ＝ 0.0636
Log likelihood ＝ −113.17539　　Pseudo R 2 ＝ 0.0150

breast	Coef.	Std. Err.	Z	P>| z |	[95% Conf.	Interval]
treat	−.671819	.3643271	−1.844	0.065	−1.385887	.0422491
_cut 1	−1.03708	.282662	(Ancillary parameters)			
_cut 2	.2156908	.2632804				

において 0 または 1 の値をとる独立変数が一つだけという時には，順序回帰モデルはマン・ホイットニー検定と等価である．ノンパラメトリック法全般に関して順序回帰モデルの利点は，回帰係数の有効推定量が得られることと，他の交絡変数も含めた解析へと拡張できる点にある．

解析上，印刷工の妻を 1 そして農夫の妻を 0 とコード化した．従属変数は 1，2，3 とコード化したが，実際の多くのパッケージでは任意の正の数すべてを許容している．コンピュータによる解析結果は**表 6.4** で与えられている．残念なことに，コンピュータの出力結果ではオッズ比は与えられないので，計算しなければならない．この結果，オッズ比は exp（−0.672）＝0.51 で，その 95% 信頼区間は exp（−1.386）～exp (0.042)，すなわち，0.25～1.04 となる．従属変数に関して 2 カテゴリーだけを有する時なら，表 3.2 で得る 0.47 のオッズ比（95% 信頼区間：0.21～1.05）と対照的である．この解釈は 1 ヵ月後と 3 ヵ月後で印刷工の妻は，農夫の妻と比べて，同じ授乳カテゴリーにあるオッズの半分かまたはより高いということになる．

尤度比 χ^2 統計量は自由度 1 で，モデルで項が 1 つであることに対応する．それに関連した P 値は 0.0636 であり，ワルト統計量の P 値 0.065

に近い．二つの切片は出力では_cut 1 と_cut 2 と出力されている．これらは**補助的パラメータ**（ancillary parameters）として知られており，モデル当てはめのために導入した追加パラメータであるが，どの要因が関与しているかといった要因の推論には関係ないし，有意かどうかということも問題にしていない．

順位データのための比例オッズモデルとその他のモデルの有益な議論は，Armstrong and Sloan[7] と Ananth and Kleinbaum[6] に与えられている．その他のモデルには，**連続比モデル**（continuation ratio model）が含まれている．Armstrong and Sloan[7] は，ロジスティックモデルの代わりに比例オッズモデルを用いたとしても有効性の上で得することがあるかといえば，たいてい大きくはないと結論づけている．順位変数を二値化しロジスティック回帰モデルを使用する戦略は，主要な予測変数の係数が有意性の境界に近くなければ，単純であるということと解釈が容易だということから十分に推奨できる．

6.2.2 モデルの検討

いくつかの検定が比例オッズモデルのために利用できるが，これらの検定は検出力が乏しい．このモデルは比例オッズ仮説からやや乖離していても頑健である．crude な検定で各分点に関連するオッズ比を吟味し，これらがすべて**単位**（unity）よりも大きいか小さいならば，比例オッズモデルを十分使用できる．表 6.3 から以下のようなオッズになる：

1 ヵ月未満対 1 ヵ月以上　　3 ヵ月未満対 3 ヵ月以上

$$\text{オッズ比}=\frac{15\times30}{20\times40}=0.56 \qquad \text{オッズ比}=\frac{30\times14}{36\times25}=0.47$$

これらのオッズ比は互いに全く近い値をとり，比例オッズモデルから 0.51 の観測オッズ比は二つのオッズ比の間にある．つまり，比例オッズモデルがあてはまらないと棄却する理由はないことになる．モデル検

定は複数の入力変数があり，ましてそれらのあるものが連続変数である時は十分に複雑であり，専門家の助けを求めるべきである．

6.2.3 順序回帰の実例

Hotopf ら[8]は，7，11，15 歳での 3 つの連続した調査で計測された慢性の小児の腹痛と 36 歳での成人の精神的な疾患との関係を 3,637 人のコホートで検討した．精神疾患の 7 点指標，すなわち『定義のインデックス（index of definition)』は結果変数として測られた．これは順位尺度である．3 調査すべてに関して痛みである二値予測変数（原因変数）は，潜在的な交絡因子である性，父親の社会的地位，36 歳での既婚/未婚，そして教育水準がモデルに取り込まれた時には，オッズ比 2.72（95% 信頼区間：1.65～4.49）で関わっていた．このようにして腹痛を伴う小児は，後生精神的な問題を発現するらしいと著者らは結論づけた．定義の指標として通常の分割は 5 であるが，全体の尺度の使用はさらに多くの情報を使用し，その結果，より精度の高い推定量を与えてくれる．

6.3 時系列回帰

時系列回帰は従属変数と独立変数が経時的に測られる状況に関したものである．通常，一つの従属変数と多くの独立変数をもつ単一の系列だけで，データのいくつかの系列がある時の反復測定とは似ていない．

時系列回帰モデルでは交絡の可能性が非常に高く，多くの変数が時間軸上で単純に増加したり減少したりして，その結果で系列的に相関をもったりする[9]．さらに，多くの疫学的な変数は季節性をもち，その変動はその因子が因果的に関連していない時でさえも現れる．季節性とその**傾向**（trend）は適切に捉えることが大切である．結果変数は季節的なので，単純に予測変数がもつ季節性ゆえの因果性に帰することが不可能なのである．例えば，幼児の突然死は夏季よりも冬季に多いが，これ

は気温が因果因子として反映しているわけではない．すなわち，日光の減少またはウイルスの存在のような多くの他の要因が影響しているかもしれない．しかし，予期しないほど寒い冬が幼児の突然死の増加に関係しているとか，非常に寒い日々が短期間一貫して訪れた後に，ごく一般的な幼児の突然死を上昇させるなら，多分因果性があると思うかもしれない．

しばしば，交絡因子が正確に捉えられたならば，残差の系列相関は現れない．すなわち，時間従属な予測変数に関連しているので系列的に相関し，その結果，残差が独立であるこの変数に関して条件付きとなる．これは，伝染病を除き個体死が相関していないところでは，死亡データにとっては特に当てはまるらしい．このようにして，残差の系列相関を検討し，そして独立性からの乖離に関する明らかな証拠がないかどうかを確認のうえ，無ければ通常の回帰手法を用いることができる．

残差の系列相関を除くために既知のまたは潜在的な交絡因子を取り込めなかった時は，通常の最小二乗法ではパラメータの標準誤差を正しく推定してこない．

6.3.1 モデル

結果変数が連続変数の場合にはモデルを次のように仮定する：

$$y_t = \beta_0 + \beta_1 X_{t1} + \ldots + \beta_p X_{tp} + \nu_t, t = 1, \ldots, n \qquad (6.5)$$

方程式（2.1）からの主な差異は，個体よりもむしろ今や t を指示することにある．同じ共変量をもつ2個体が交換可能であるにもかかわらず，例えば日曜日と土曜日を交換はできず同じ結果を期待できないので，時点を区別することが大切である．誤差項を ν_t とし，$\nu_t = \varepsilon_t - \alpha\nu_{t-1}$ と仮定する．ここで，ε_t は平均0，分散 σ^2 で独立に正規分布する変数であり，α は -1〜$+1$ の値を取る定数と仮定する．誤差項は（1次の）**自己回帰過程**（autoregressive process）として知られている．このモデルはデータが時間とともに相関しており，**系列相関**（serial correlation）とし

て知られている．系列相関を無視すると回帰係数の標準誤差を人工的に低く推定する効果がある．このため関連がないという帰無仮説の下で提案されている有意水準よりも，しばしば大きな水準となってしまう．

6.3.2 相関のある残差を用いた推定

上述のモデルを与え，α が既知であると仮定すると，Cochrane–Orcutt 手順[9] として知られている一般化最小二乗法を使うことができる．

単純にするために，独立変数は1つと仮定し，$y_t^* = y_t - \alpha y_{t-1}$，$x_t^* = X_t - \alpha X_{t-1}$ と書く．それから，y_t^* と x_t^* について通常の最小二乗法を用いて β の推定量を求めることができる．しかし，α は通常未知なので，

$$\alpha = \sum_{t=2}^{n} e_t e_{t-1} \Big/ \sum_{t=2}^{n} e_{t-1}^2$$

により通常の最小二乗残差 e_t から推定できる．

このことから，変換された変数の集合と回帰推定量の集合を組み立て，以下収束するまで続けるといった繰返し手順を用いる．この繰返し Cochrane–Orcutt 手順は，最初の個体 y_1 を固定とみなしたところで，α と β の最尤推定量を計算するステップワイズ・アルゴリズムと解釈できる．残差が正規分布すると仮定される時は Full の最尤法が利用できて，α と β を同時に推定する．この方式は高次の自己回帰モデルに一般化され，特に SAS のような多くのコンピュータパッケージで実行できる．しかし自己相関が高い時に，この手法を用いるのは注意が必要である．自己回帰誤差モデルはよく当てはまらない時の代案として使うべきではない，と指摘しておきたい．

最初の点が固定すると仮定しないモダンな手法もあるが，データセットが長い（例えば，50 ポイントより大）場合にはあまり効果がない．これらのモデルは頻度である結果変数へと一般化されるが，本書の範囲を超えてしまう．より詳しくは，Campbell[12] を参照のこと．

6.3.3 コンピュータ出力の解釈

表 2.1 の死腔と身長に関するデータは，実際は時間軸上で追跡される 1 人の個体のデータであるとする．それから，身長に対する死腔の回帰は Cochrane–Orcutt 回帰を用いて表 6.5 で与えられる．この手法は最初の観測値を無駄にし，そのため回帰係数は図 2.1 のものと厳密には比較できない．この出力は観測数が 15 ではなく 14 とされていることに注意しよう．ここで得られた標準誤差 0.214 は，各点がすべて独立だと仮定した時の 0.180 よりも十分大きい．自己相関係数 α の推定量はこの出力では rho と表わされ，0.046 とかなり小さい．しかしこのプログラムでは rho に対する P 値を与えてはいない．

表 6.5　1 個体に属する時点がすべて時間軸上で均等に配置されていると仮定した場合の，表 2.1 のデータに対する Cochrane-Orcutt 回帰分析の結果

```
Iteration 0:   rho= 0.0000
Iteration 1:   rho= 0.0432
Iteration 2:   rho= 0.0462
Iteration 3:   rho= 0.0463
Iteration 4:   rho= 0.0463
Iteration 5:   rho= 0.0463

Cochrane-Orcutt AR (1) regression iterated estimates
```

Source	SS	df	MS	Number of obs	=	14
Model	4841.31415	1	4841.31415	F(1,12)	=	29.29
Residual	1983.76032	12	165.31336	Prob>F	=	0.0002
Total	6825.07447	13	525.005728	R-squared	=	0.7093
				Adj R-squared	=	0.6851
				Root MSE	=	12.857

Deadspce	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Height	1.160173	.2143853	5.412	0.000	.6930675	1.627279
_cons	−102.1168	31.78251	−3.213	0.007	−171.3649	−32.86861
rho	.0463493					

Durbin-Watson statistic (original) 1.834073
Durbin-Watson statistic (transformed) 1.901575

6.4 ポアッソン回帰，順序回帰，時系列回帰について，文献での報告の仕方

- 独立変数が離散的な量的変数ならば，ポアッソン回帰は必要とされるモデルかもしれない．モデルの当てはまりの良さを示す χ^2 統計量を引用して，そのモデルがデータに当てはまっていることを証明しておこう．モデルがよく当てはまっていない時には，共変量の交互作用を検定するか，または過剰ポアッソン変動を許容してみよう．
- 従属変数が順位である時は順序回帰が役に立つかもしれない．しかし順位変数のカテゴリー数が大きい（例えば7以上）時には，線形回帰のほうが適当かもしれない．多分主要な独立変数のための各分点に関連したオッズ比を引用することにより，比例オッズモデルは適当なモデルであることを見ておこう．比例オッズが当てはまりそうもないなら，従属変数を二値化して，ロジスティック回帰を使ってみよう．主要な独立変数のために一番有意性を与えるものを選び，二値化のための分点を選んではいけない．
- データが時系列を形成する時，モデルの残差が系列的に相関している証拠を探しなさい．その証拠があるならば，系列相関を許容するような項をモデルに取り込みなさい．

6.5 ポアッソン回帰，順序回帰，時系列回帰の結果について，文献から読み取れること

- 通常，モデルが適切である証拠を探しなさい．
- ポアッソン回帰では頻度は独立か？　そうでないのなら過剰分散が考えられるか？
- 順序回帰を使ってきたのならどのように結果が解釈されたのか？
- 時系列回帰での通常の誤りは系列相関を無視することである．このこ

とにより解析が無効になることはないかもしれないが，そうなのかどうかを尋ねてみる価値はある．別のよくある特徴は系列相関のために許容される1次の順序自己回帰を使用することだけであるが，それが十分かどうかを尋ねる価値はある．

■文献

1 McNeil D. *Epidemiological Research Methods*. Chichester: John Wiley, 1996.
2 Breslow NE, Day NE. *Statistical Methods in Cancer Research: Vol II – the design and analysis of cohort studies*. Lyon: IARC, 1987.
3 Doll R, Hill AB. Mortality of British doctors in relation to smoking: observations on coronary thrombosis. *Nat Cancer Inst Monog* 1996; **19**: 205–68.
4 Campbell MJ, Cogman GR, Holgate ST, Johnston SL. Age specific trends in asthma mortality in England & Wales 1983–1995: results of an observational study. *BMJ* 1997; **314**: 1439–41.
5 Swinscow TDV. *Statistics at Square One, 9th edn* (revised by MJ Campbell). London: BMJ Books, 1996.
6 Armstrong BG, Sloan M. Ordinal regression models for epidemiologic data. *Am J Epidemiol* 1989; **129**: 191–204.
7 Ananth CV, Kleinbaum DG. Regression models for ordinal responses: a review of methods and applications. *Int J Epidemiol* 1997; **26**: 1323–33.
8 Hotopf M, Carr S, Magou R, Wadsworth M, Wessely S. Why do children have chronic abdominal pain and what happens to them when they grow up? Population based cohort study. *BMJ* 1998; **316**: 1196–200.
9 Yule GU. Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time-series. *J Roy Statist Soc* 1926; **89**: 187–227.
10 Cochrane D, Orcutt GH. Application of least squares regression to relationships containing autocorrelated error terms. *J Am Statist Assoc* 1949; **44**: 32–61.
11 SAS Institute Inc. *SAS/ETS User's Guide Version 5 Edition*. Cary, NC: SAS Institute, 1984: 192.

12 Campbell MJ. Time series regression. In: Armitage P, Cotton T, eds. *Encyclopaedia of Biostatistics.* Chichester: John Wiley, 1997: 4536–8.

付録 1：指数と対数

対　数

ある値のベキ乗を取ることは簡単である．例えば，$y=x^2$ というのは $y=x.x$ のことである．一般に任意の n に対して，$y=x^n$ とは $y=x.x...x$（n 回掛ける）のことである．

ここで次のことがすぐわかる．任意の n，m に対して

$$x^n.x^m = x^{n+m} \qquad \text{(A 1.1)}$$

である．したがって，$3^2\times3^4\times3^6=729$ となる．この公式は数値でなくても，いかなる値（n と m）に対して成立つ．

ここで $x^0=1$ と定義する．その理由は，$x^n = x^{0+n} = x^0x^n = 1.x^n$ となるためである．

ベキ乗の考え方を少し拡張すると便利である．すなわち，ベキ乗として分数や負の値を使えるようにするのである．例えば，$y=x^{0.5}$ というのは $y=\sqrt{x}$ に等しいとする．その理由だが，$x^{0.5}.x^{0.5} = x^{0.5+0.5} = x^1 = x$ となるからである．いいかえると $\sqrt{x}\cdot\sqrt{x}=x$ となる．

同様にして x^{-1} は $1/x$ に等しくなる．なぜなら $x.x^{-1} = x^{1-1} = x^0 = 1$ になるからである．

もし $y=x^n$ であるとすると，y に関する x の対数が定義できて，通常 $n=\log_x(y)$ と書く．そして，「底 x での y の対数は n である」という．

ここで $y=x^n$，$z=x^m$ としよう．（A 1.1）式より次のことが分かる．

$$\log_x(y.z) = n+m = \log_x(y)+\log_x(z)$$

このことから，2つの数字を掛けるということは，それらの対数を足すことになる．これこそが，対数を使う理由の大きなものである．なぜなら対数を用いると，掛け算をする代わりに足し算で済み，手計算で簡

単にできるからである．［付録 2］でこれと同じような結果を用いるが，それは以下のとおりである．

$$\log_x(y/z) = \log_x(y) - \log_x(z)$$

いいかえると，2 つの数字の比の対数を取ると，それは 2 つの対数の差になる．

ここで最も通常の底とは 10 であり，ちょっと変わった底としては e =2.718...がある．点々は，小数点が無限につづくという意味である．この数値を使うと $y = e^x$ という曲線の傾きは，どのポイント (x,y) でもちょうど y という性質がある(訳注)．ところが，他の底を用いると傾きは y に比例はするものの，y 自身に等しくなるとは限らない．なお，$y = e^x$ という式は $y = \exp(x)$ と書くこともある．底が e または 10 の対数は，電卓上では ln（e のほう）または log（10 のほう）で示される．前者のほうを**自然対数**（natural logarithm）とよぶ．この本ではすべての対数が自然対数，つまり底 e である．ある底から別の底に変換することもできる．つまり，$\log 10\, y = \log_e y . \log_{10} e$ として底 10 から底 e に変換できる．e はどんな値かを電卓で確かめるには，1 を入力して exp を押せばよい．$\log_{10}(e)$ は 0.4343 という定数である．したがって，$\log_{10} y = 0.4343 \times \log_e y$ となる．

それでは電卓で計算してみよう．何か正の数字を入力して，ln を押してみてください．そのあと exp を押してください．元の数に戻ったことに気づきますね．その理由は $\exp(\ln(x)) = x$ だからである．

0 より大きな数字 x が底であれば，定義から $\log_x(1) = 0$ となる．それでは，電卓で $\log_x(1)$ と $\ln(1)$ を計算してみてください．

この本で，指数と対数というのは何度も出てくる．データをモデルに当てはめるとき，線形予測子として足し算の形で表したほうが分かりやすい．しかし，例えば第 3 章に示したリスクのように掛け算で示されるモデルもある．このとき対数を取れば，掛け算は足し算になるので便利

訳注：$y = e^x$ を微分して傾きを求めると e^x そのままだから．

である．対数はさらに次の用途にも使う．それは**右に歪んだ**（positively skewed）分布を変換するときに用いる．そうすることで，正規分布に近い分布になることが多いためである．もちろん，ゼロや負の値があるとこれはうまくいかない．

付録 2：最尤法と有意性検定

要　旨

この付録では，**最尤法**（maximum　likelihood）の使途について簡単に説明する．この方法は，本書で述べてきたモデル当てはめに用いるものである．まず**ワルト検定**（Wald test）と**尤度比検定**（likelihood ratio test）について説明し，次に**乖離**（deviance）との関係を示す．もっと詳しいことは，Clayton and Hill[1] を参照のこと．

二項モデルと尤度

モデル（model）とはデータを記述するための構造のことであり，2つの部分から成っている．第1の部分では，説明変数が予測変数と線形でどのように関連するかを記述する．そのあと**リンク**（link）関数というもので変換され，結果変数に関する予測値あるいは当てはめ値が得られる．第2の部分では，結果変数に関する予測値の周りの確率分布を示す．

多分一番シンプルなモデルが二項分布である．イベントは確率 π で生じる．例えば，男の子が生まれることをイベントと仮定し，5人の母親は2人の男の子と3人の女の子を生むと仮定しよう．男の子は1番目と3番目に生まれた．もし男の子である確率を π とすると，このようなことが生じる確率は，$\pi \times (1-\pi) \times \pi \times (1-\pi) \times (1-\pi)$ になる．もし母親の年齢などが異なっているなら，それらを区別したいだろう．i 番目の母親が男の子を生む確率を π_i と書けば，女の子を生む確率は $(1-\pi_i)$ となる．そこで，$\pi_1 \times (1-\pi_2) \times \pi_3 \times (1-\pi_4) \times (1-\pi_5)$ となる．内

容を表す少し見慣れぬ用語だが，この確率のことを**尤度**（likelihood）とよぶ．『尤度（likelihood)』も『確率（probability)』も通常では同じ意味である．これを$L(\pi)$と書くが，尤度とはモデルが**指定された**（given）ときの，そのデータが現れる確率のことである．

最尤推定（maximum likelihood）の考え方とは，尤度を最大にするようなπを決めることである．第3章で，私たちはπのモデルについて議論した．このπは，個人の背景因子の関数である．確認のため，次に2つの場合を考えてみよう．πは共通であり，個人を区別する情報はない場合が第1である．もう一方は各πはデータによって決まり，結果が男の子か女の子も各πで決まる場合である．後者の場合，$\pi_1=\pi_3=1$で$\pi_2=\pi_4=\pi_5=0$を選ぶ．これを**飽和モデル**（saturated model）とよぶ．なぜなら，モデルがパラメータで飽和しているからである(訳注)．最大パラメータ数とは，データ数〔厳密にいうと**自由度**（degrees of freedom)〕になる．この場合，尤度は

$$L(\pi)=1\times(1-0)\times1\times(1-0)\times(1-0)=1$$

となる．

πがすべて共通だとすれば，$L(\pi)=\pi\times(1-\pi)\times\pi\times(1-\pi)\times(1-\pi)=\pi^2(1-\pi)^3$となる．一般的に，$N$人の子供のうち$D$人が男の子であれば，$L(\pi)=\pi^D(1-\pi)^{N-D}$となる．$\pi$がいかなる数値を取っても，この尤度の値は大変小さくなるので，尤度よりも対数尤度のほうが便利である．すなわち，

$$\log(L(\pi))=D\log_{(\pi)}+(N-D)\log(1-\pi)$$

となる．

尤度を最大にするπは，対数尤度も最大にすることは自明である．

この式で，NとDはデータから与えられる．ここで統計的問題とは，πが変わると$\log(L\pi)$はどう変化するかを見ることになる．そして，デー

訳注：5人の母親がそれぞれパラメータをもつので，データ数5人と同じだけパラメータが存在するため．

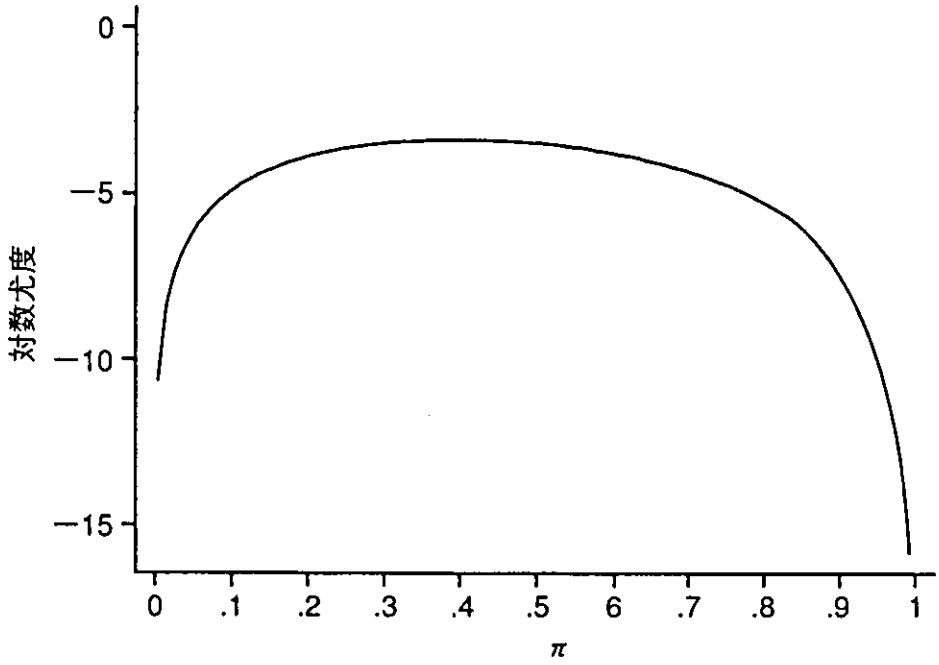

図 A 2.1　D=2，N=5 の二項分布における π に対する対数尤度のグラフ

タに最も適合する π を選ぶ，それが log(Lπ) を最大にする値 π になる．上のデータ（2 人の男の子と 3 人の女の子）で対数尤度のグラフを書くと，図 A 2.1 のようになる．

最大値は π＝0.4 のところに見られる．これは常識でも分かるところである．最大での対数尤度値は

$$\log(L(\pi_{max}))=2\log(0.4)+3\log(1-0.4)=-3.3651$$

である．

グラフは大変フラットであるが，それは最大値が推定しにくいことを意味している．たった 5 人の限られた情報のためである．

理由は後で述べるが，尤度は最大尤度値で標準化したもの，つまり尤度比 LR(π) に書き直すことも多い(訳注)．すなわち

$$LR(\pi)=L(\pi)/L(\pi_{max}) \quad 0<\pi<1$$

である．

訳注：Likelihood Ratio の略で LR と書く．

この対数（[付録 1] 参照）を取ると

$$\log LR(\pi) = \log\{L(\pi)\} - \log\{L(\pi_{max})\}$$

となる．

この最大値は同じく $\pi = 0.4$ のときであるが，この場合最大値はゼロとなる．

ポアッソンモデル

第 6 章で述べたポアッソンモデルは，対象数 N が大きく，イベント確率 π が小さいときに有用である．すなわち，イベントの期待数 $\lambda = N\pi$ が中程度になる．

このとき対数尤度は

$$\log(L(\lambda)) = D\log(\lambda) - \lambda$$

になる．

正規モデル

平均 μ，標準偏差 σ の正規分布に従う変数 Y の確率分布は，次のように与えられる．

$$\frac{0.3989}{\sigma}\exp\left[\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right]$$

この値は μ と σ が異なれば変化する．もし σ が既知（さらに固定）ならば，尤度は上に示した確率になるが，σ によっては変化しなくなる．そして，それは μ だけの関数となるので $L(\mu)$ と書く．

大人男性の身長は正規分布すると仮定しよう．私たちは平均 μ については知らないが，標準偏差は 15 cm だと知っているとする．この母集団からランダムに 1 人の男性を選んだところ，その身長は 175 cm であった．このとき，このデータに関する対数尤度は以下のとおりである．