ベントが起こる人は何%いたかを計算することができる．例えば，ある疾病の有無を性別（2カテゴリー）と社会階層（5カテゴリー）で見たいとしよう．そうすると，性別と社会階層から10カテゴリーの表ができ，各グループで疾病の割合を検討できる．これが第1の状況である[訳注1]．

2．第2の状況は人数分の行からなる表があり，疾病の割合としては0か1しか入らない場合である．独立変数の少なくとも1つが連続変数であると，こういうことになる．これはデータを入力するときの方法でもある．各人は別々であり，決してグループ化したくはないだろう[訳注2]．

もし表形式のデータなら，コンピュータプログラムごとにコマンドが異なる．個人別のデータでも同じ回帰推定値が得られるが，こちらではもっと流動的な解析が可能となる．このことについては，3.3節でもっと詳しく議論されている．

Swinscow本[1]を思い出そう．統計解析というのは，**標本**（samples）から**母集団パラメータ**（population parameters）を推定するものである．ロジスティック回帰では母集団パラメータをモデル化する．まず，カテゴリー別でグループ化したほうを考える．セルiのイベント発生母集団確率をπ_iと置く．これはまた『期待（expected)』値でもある．もし偏っていないコインだったら，『表（head)』が出る母集団あるいは期待確率は0.5である．従属変数y_iはイベント発生割合ということになる（何回かのコイン投げで表の出る割合のこと）．このとき，$E(y_i)=\pi_i$と書き，ここでEは『期待値（expected value)』を意味する．さらに，イベントは確率π_iで起こるとしたので，イベントの**オッズ**（odds）は$\pi_i:(1-\pi_i)$，つまり$\pi_i/(1-\pi_i)$になる．コイン投げの場合，表の裏に対するオッズは1：1(=1）になる．

訳注1：表形式データのこと．
2：個人別データのこと．

そこでこのモデルは

$$\log_e\{\pi_i/(1-\pi_i)\} = \mathrm{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}. \qquad (3.1)$$

と書ける．ここで，独立変数は $X_{i1},\ldots,X_{ip}$ である．

この式の左辺は成功の対数オッズであり，しばしばこれを**ロジスティック** (logistic) 変換あるいは**ロジット** (logit) 変換と呼ぶ．

モデル (3.1) がなぜ良いかというと，係数 β は2×2表の**オッズ比** (odds ratio) と関係しているからである．今，**共変量** (covariate) が一つしかない場合を考えよう．しかも二値で0か1しか値を取らないとする．このとき，x と y の関連を示す**オッズ比** (odds ratio) は $\exp(\beta)$ になる．時々オッズ比であっても相対リスクと呼ぶようだが，このオッズ比は『相対リスク (relative risk)』とは**異なる**ことに注意しよう．もし x が連続変数だとしたら，$\exp(\beta)$ は x の単位増分に伴うオッズ比ということになる．

ロジット変換を正当化するには，第1にオッズ比は二値変数を扱うのに自然なパラメータであること，第2にロジット変換によりオッズ比と独立変数を関連づけるのが容易なためである．次のようにも正当化できる．(3.1) 式の右辺は $-\infty$ から $+\infty$ までの値を取りうる．左辺はというと，確率は0から1なのでオッズ比は0から ∞ となり，ロジットを取ると $-\infty$ から $+\infty$ になり，右辺と同じになるためである．

この段階では，従属変数の**観測** (observed) 値は式に入っていない．それは二項分布（[付録2] 参照）のモデル化として関係づけている．セル i において n_i 人中 y_i の成功を観察するとき，y_i は確率 π_i の二項分布だと仮定する．モデル中のパラメータは最尤法で推定される．この方法についても [付録2] を見てもらいたい．当然，私たちは母集団の値 π_i については知らない．モデリングの過程では，その推定値あるいは当てはめ値をモデルに代入する．

重回帰との類推のため，このモデルは上に示したように記述することが多い．しかし，ここでは観測した割合 p_i ではなく π_i に置き換わって

いる．このためモデルから両者をつなげる誤差分布が抜けている．実際には，観測された割合をモデル当てはめして，重回帰のように最小二乗法が使える．このとき p_i が0か1に近いと，あまり当てはめはよくない．しかし，このモデルの解釈は（3.1）式の解釈とは異なる．オッズ比とのリンクがないからである．現代のコンピュータを用いれば最尤法で簡単に求められるし，しかもこちらの方法のほうがよい．従属変数が0または1だとロジットが存在しない．こうなるとロジスティック回帰は不可能と思う人もいるだろう．しかし，先に説明したように，このモデルでは**期待**（expected）値のロジットを用い，観測値のロジットは用いていない．このモデルでは期待値が0より大きくて，1よりも小さいことは保障している．

私たちは，イベントの起こる確率を計算したいかもしれない．そのためには（3.1）式の係数が b_0，$b_1,\ldots,b_p$ と推定されたと仮定する．第1章から線型予測子の測定値は

$$\mathrm{LP}_i = b_0 + b_1 X_{i1} + \ldots + b_p X_{ip}$$

と表せる．

式（3.1）は，また次のように書ける．

$$\hat{\pi}_i = \frac{e^{\mathrm{LP}_i}}{1 + e^{\mathrm{LP}_i}} \qquad (3.2)$$

ここで $\hat{\pi}_i$ は π_i の推定値である．この式を用いると，モデルから確率を推定できる．これらは，y_i の予測値あるいは当てはめ値に等しい．モデルがよく当てはまっていれば，観測割合 y_i/n_i は π_i に近づくだろう．

ロジスティック回帰のさらなる記述は，Collett[2] と Hosmer and Lemeshow[3] に見られる．

3.2 ロジスティック回帰の用途

1．断面調査，コホート調査，臨床試験で結果変数が二値のとき，重回

帰分析の代わりに用いる．ロジスティック回帰を用いると，原因変数と二値の結果変数の関連性を研究することができる．そのときに交絡変数はいくつか含んでいてもよく，それらはカテゴリー変数でも連続変数でもあってもよい．

2．判別分析のように，二つの群を判別するための因子を見つけたいときに用いる．ここでの結果変数は，どちらのグループに入っているかを示す二値変数になる．一例を挙げると，心理テストの結果から男女を判別したいようなときがあろう．

3．外科手術後の合併症を発症するリスクのように，予後因子を確立したいとき．

4．ケースコントロール調査や，マッチングを伴うケースコントロール調査を解析するとき．

3.3 コンピュータ出力を理解する:グループ化した解析の場合

ほとんどのコンピュータプログラムでは，データがグループ化されているか，それとも個人別になっているかによって別のプロシージャを使う．個人ごとにデータを保存しておくほうが便利なのは，いろいろな目的にそれを使えるからである．しかし仮に独立変数(訳注)がすべてカテゴリーならば，ロジスティック回帰分析による係数とSEは，グループ化された方法を用いたときと等しくなる．グループ化したほうでは，従属変数はそのグループ内での成功の回数となり，グループ化しない場合の従属変数は単に0か1になる．一般的に言って，モデル適合度をみるにはグループ化したほうが容易である．

Swinscow本[1]の中の例を取り上げる．生後3ヵ月で母乳と哺乳瓶の違いと，そのお母さんが印刷屋の妻か農家の妻の違いの関連性を見ようとした．赤ちゃんが3ヵ月以上母乳を与えられたらイベントと定義し，母

訳注：原文は従属変数とあるが訂正した．

表 3.1　赤ちゃんに授乳を 3 ヵ月未満または 3 ヵ月以上した妻の印刷屋と農家の人数
(Swinscow 表 8.3 より引用)

妻	3 ヵ月未満	3 ヵ月以上	合計
印刷屋の妻	36	14	50
農家の妻	30	25	55
合計	66	39	105

グループ化した解析用としてこのように書ける

Y	n	職業 (1＝印刷屋，0＝農家)
36	50	1
30	55	0

グループ化しない解析用として次のようにも書ける

Y (授乳) (1＝3 ヵ月未満， 0＝3 ヵ月以上)	職業 (1＝印刷屋，0＝農家)
1 (36 回)	1
1 (30 回)	0
0 (14 回)	1
0 (25 回)	0

親が農家の妻なら 1，印刷屋の妻なら 0 とするカテゴリー変数 X_1 だけを設けた．データは表 3.1 に示したとおりである．

これらのデータ (2 番目の個人別のほう) でロジスティック回帰を行った出力が，表 3.2 である．前半は定数項についての結果であり，後半は職業の項についてである．出力にはモデルの係数を示すようにもできるし，オッズ比を示すようにも設定できる．もちろん，両方とも出力することも可能である．出力には対数尤度値も示されているが，これについ

表 3.2　表 3.1 のデータを用いてロジスティック回帰分析した結果

Logit estimates　　Number of obs = 105
LR chi 2(0) = 0.00
Prob > chi 2 =
Log likelihood=−69.26972　　Pseudo R2 = 0.0000

_outcome	Coef.	Std. Err.	z	P>| z |	[95% Conf.	Interval]
_cons	.5260931	.2019716	2.605	0.009	.130236	.9219502

Logit estimates　　Number of obs = 105
LR chi 2(1) = 3.45
Prob > chi 2 = 0.0631
Log likelihood=−67.543174　　Pseudo R2 = 0.0249

breast	Coef.	Std. Err.	z	P>| z |	[95% Conf.	Interval]
Occupatn	−.7621401	.415379	−1.835	0.067	−1.576268	.0519878
_cons	.9444616	.3149704	2.999	0.003	.327131	1.561792

breast	Odds Ratio	Std. Err.	z	P>| z |	[95% Conf.	Interval]
Occupatn	.4666667	.1938435	−1.835	0.067	.2067453	1.053363

ては［付録 2］で述べている．ある種の**平方和**（sum of square）と考えられる．

職業項なしのモデルでの対数尤度は−69.27 であり，職業項を含むと−67.54 になる．その差に(−2)を掛けると，つまり，「LR chi 2(1)」のところ（LR とは尤度比 likelihood ratio）が 3.45 と表されている．これは自由度 1 の χ^2 分布に従っている．詳しくは［付録 2］を参照のこと．このモデルの尤度比 χ^2 統計量の自由度が 1 なのは，モデルには**職業**（occupation）という項しか含んでいないためである．ちなみに，職業項は非有意である（P=0.0631）．

擬似 R 2 値（Pseudo R 2」）については［付録 2］に述べられているが，対数尤度の減少に比例する量である．直線回帰のときの R^2 と同様であり，そのモデルで説明される変動の割合を示す．二値変数の割合は少し

分かりにくいので，擬似R２のおよその大きさだけ見ればよい．この例の場合0.0249であるから，そのモデルはあまりよく当てはまっていないと分かる．なぜなら，変動のごく一部しか説明されていないからである．

ワルト統計量（Wald statistic）とは推定値bをそのSEで割った値であるので，$z = b/\mathrm{SE}(b) = (-0.7621/0.4154) = -1.85$となる．それを二乗すると3.3665であり，尤度比統計量と近いことがわかる．また，P値（0.067）も尤度比χ^2に近い．

Swinscow本[1]に示されているような従来のχ^2統計量というのは，ワルトでも尤度比でもない．いわば3番目の統計量であり，尤度論から導かれる**スコア統計量**（score statistic）という．［付録2］を参照のこと．この値は3.418で，他の２つの統計量に近い．

もしb_1がβ_1の推定値なら，$\exp(b_1)$がX_1に対する推定オッズ比になる．表3.1からオッズ比を求めると，(30×14)/(36×25)＝0.4667と計算される．これは表3.2の出力にも表れている．このモデルの係数は−0.7621なので，オッズ比（OR）は$\exp(-0.7621) = 0.4667$と計算できる．印刷屋の妻は農家の妻に比べて，母乳を与えるのは半分位と分かる．

β（つまり対数オッズ）の95%信頼区間は，b±1.96×SE(b)で与えられる．これは**ワルト信頼区間**（Wald confidence interval）（3.4節を見よ）といわれる．ワルト検定に基づいているからである．こうしてオッズ比の95%信頼区間は，$\exp\{b-1.96\times\mathrm{SE}(b)\}$から$\exp\{b+1.96\times\mathrm{SE}(b)\}$になる．これはオッズ比（OR）に関して対称ではない．直線回帰のときの信頼区間とは違うのである．一例を挙げると，表3.2から信頼区間を計算すると，$\exp(-0.7621-1.96\times0.4154)$から$\exp(-0.7621+1.96\times0.4154)$なので0.207～1.053となり，これは0.4667の周りに対称ではない．信頼区間が1を含むということは，有意性検定において5%水準で非有意を意味している．このことは一般的には正しいが，検定と少し食い違う結果になることもある．それはオッズ比が大きいとき

である．その理由は，有意性検定は尤度比検定かスコア検定に基づいているが，信頼区間はワルト検定に通常基づいているためである．

3.4 ロジスティック回帰の実際

Lavie ら[4]は，睡眠障害の疑いありとして睡眠外来に照会された 2,677 人の大人を調査した．彼らは睡眠障害の重症度を定義し，高血圧の有無との関連性を調べた．

彼らにとっての質問は以下のとおりである．

(i) 年齢，性，BMI (body mass index) で補正したうえで，障害スケールは高血圧の予測因子になっているのか？

(ii) 他の共変量で補正したうえで，性別は高血圧の予測因子になっているのか？

表 3.3 に結果を示したが，回帰係数（対数オッズ）とワルト信頼区間である．ダミー変数である性別に関する係数は 0.161 なので，男性が高血圧をもつオッズは，exp(0.161) = 1.17 倍女性よりも高いことが分かる．オッズ比の 95% 信頼区間は exp(−0.061) から exp(0.383)，すなわち 0.94〜1.47 になる．これは 1 を含んでいるので（回帰係数の信頼区間でいうとゼロを含むとなる），性別は高血圧の有意な予測因子とはいえない．年齢についてはどうであろうか．性別が同じ 2 人がいて，しかも BMI も同じであり，年齢は一方が他方より 10 歳年上だと仮定しよう．このとき，一方の人は他方より 2.24 倍高血圧になりやすいことにな

表 3.3 高血圧のリスクファクター[4]

リクスファクター	推定値（対数オッズ）	（ワルト 95% 信頼区間）	オッズ比
年齢 (10 歳)	0.805	(0.718〜0.892)	2.24
性別 (男性)	0.161	(−0.061〜0.383)	1.17
BMI (5 kg/m^2)	0.332	(0.256〜0.409)	1.39
睡眠障害指数 (10 単位)	0.116	(0.075〜0.156)	1.12

る．10歳を選択した理由は，年齢の区分がそうなっているからである．ここで注意しておかないといけないことは，オッズ比の下では**乗法的**（multiplicative）になっているが，対数オッズでは**加法的**（additive）になっていることである．したがって，ある女性よりも10歳年上の男性では，2.24×1.17＝2.62倍高血圧になりやすい．このモデルでは，年齢と性別は高血圧に対して独立に影響しているので，それぞれのリスクを掛け算するだけでよい．独立因子かどうかは，年齢と性別による**交互作用項**（interaction term）をモデルに含めてチェックすることができる．第2章を参照のこと．これがもし有意であれば，年齢と性別のあいだには**効果修飾**（effect modification）があることを意味する．もしこの交互作用が**肯定的**（positive）ならば，年齢と性別それぞれから予測された高血圧リスクよりも，高齢男性のリスクはもっと高くなることを意味している．

3.5 モデルの確認

データをうまく表したモデルかどうかを確認する方法は色々ある．そのいくつかは2.7節の直線回帰で述べたのと同じ方法であり，予測因子の係数からその線形性をみること，影響データ，重要な交絡の欠落などを見ることである．ある個人のデータを除いてみて回帰係数がどう変化するかは簡単に見ることができるし，除くことで係数に大きな影響があるような個人を探すのも大切なことである．このような影響を与えるデータについては，重回帰と同様にロジスティック回帰でも扱える．コンピュータプログラムによっては，ロジスティック回帰でも**影響度**（measures of influence）が与えられる．

残差を定義することはロジスティック回帰では難しいし，モデル確認は直線回帰の場合とは異なる．結果変数が0または1の場合には，外れ値をチェックするのは難しい．詳細はCollett[2]とCampbell[5]に見られる．

ロジスティック回帰で特に問題となるのは次の三点である．すなわち，適合度，『超二項（extra-Binominal）変動』(訳注1)，ロジット変換，の三つである．

適合度

独立変数がすべてカテゴリーなら，各セル内での観測割合とモデルによる予測割合を比較してみることができる．しかし，一部連続変数が含まれていると，予測値を何らかの方法でグループ化しないといけない．Hosmer and Lemeshow[3] はいくつかの方法を提案している．一つの方法はこうである．モデル π_i を使った予測確率でもって 10 個のグループ（つまり 10% ずつ）に分け，グループごとに成功数の予測値を計算する．それは，そのグループの個人 1 人ずつにつき成功の予測確率を求め，それらの和で定義する．そこで観察された成功・失敗の人数と比較する際，第 5 章で述べた自由度 8 の χ^2 分布を使えばよい(訳注2)．よく当てはまったモデルであれば，各グループにつき観察された成功と失敗の人数をまずまず正しく予測できるはずである．もし χ^2 値が有意であると，そのモデルはデータをうまく表していないことになる．

『超二項』変動

重回帰では残差分散の大きさが事前には決まっておらずデータから推定されるが，ロジスティック回帰では二項分布であることから，残差分散は事前に決まっている．しかし，モデル上のパラメータは観察されたよりも小さい（時には大きい）ことがあるかもしれない．このようなとき『超二項変動（extra-Binomial variation）』と呼んでいる．分散が期待値よりも大きいときには，『過剰分散（overdispersion）』と呼んでいる．

訳注 1：二項分布で説明できない部分のこと．

2：自由度 8 なのは 10 グループからロジスティック回帰のパラメータ 2 つを引いたため．

データが独立でないときに，このようなことが起こりやすい．個人内で反復して結果を見るようなとき，開業医（GP）別の患者データなどがそうである．このようなときには回帰係数の推定値は過度に影響されるが，SE はふつう低めに推定されるため，信頼区間は狭くなりがちである．過去には近似法で対処してきた．すなわち SE は少し大きくするようにし，係数推定値はそのままにする．しかしその後，これは**変量効果モデル**（random effects model）の特殊なケースに当たることが分かり，回帰係数 β は固定ではなく，平均と分数をもった確率変数とみなす．これについては，第 5 章を参照のこと．

ロジスティック変換は不適切

確率 0 から 1 を可能性として $-\infty$ から $+\infty$ まで取りうる変数に変換するだけなら，必ずしもロジスティック変換でなくてもよい．その例としては**プロビット**（probit）変換，**2 重対数**（complementary log–log）変換，すなわち $\log(-\log(1-\pi))$ というのもある．後者はコホート調査で，死亡までのイベントがいつ起こるかを見るときに有用である．すなわち**生存時間解析**（survival analyses）に該当する（第 4 章参照）．プログラムによっては異なるリンク関数を使えるソフトもあるが，たいていは同じような結果になる．ロジスティックというリンクは分かりやすいため，一般的に推奨されているのである．

3.6 コンピュータ出力を理解する：グループ化していないデータ

連続的に 170 名の患者について，腹部手術後の合併症リスクが APACHE リスクスコア（C. Johnson から個人的に得た）で付けられた．同時に体重（kg）も測定された．このとき，結果変数は術後合併症が軽度か重度かであった．出力は**表 3.4** のとおりである．ここでは係数はオッズ比として示されている．このモデルをどう解釈すればよいか．これは

表 3.4　腹部手術データのロジスティック回帰分析の結果

(Johnston から個人的に得た)

Logit estimates　　Number of obs = 170
LR chi 2(2) = 107.01
Prob > chi 2 = 0.0000
Log likelihood = −56.866612　　Pseudo R2 = 0.4848

severity	Odds Ratio	Std, Err.	z	P>| z |	[95% Conf, Interval]	
apache	1.898479	.2008133	6.060	0.000	1.543012	2.335836
weight	1.039551	.0148739	2.711	0.007	1.010804	1.069116

Logistic model for severity, goodness of fit test
(Table collapsed on quantiles of estimated probabilities)

number of obserbations = 170
number of groups = 10
Hosmer-Limeshow chi 2(8) = 4.94
Prob > chi 2 = 0.7639

固定重みなので (for a fixed weight)，APACHE スコアが 1 単位高くなると重度合併症のオッズ比が 1.9 に増加し，それは高度に有意である（P<0.001）ということになる．

モデルの適合度で説明した Hosmer–Lemeshow 統計量は非有意であるから，観察値とモデル予測値とはよく合っている．したがって，モデルはデータによく当てはまっていることになる．実際には，Hosmer–Lemeshow 統計量でもってモデルの良さを確認し，そうであれば係数を解釈できる．しかし，人によっては係数の有意性をみる前に，適合度をみる有意性検定をするというのは間違いだとする人もいる．最初の検定[訳注]が非有意なら，モデルは正しいと言っているわけではない．単に，不適として棄却するほどの証拠はないと言っているだけである．どのモデルも完全に正しいわけでないので，データがたくさんあれば適合度検定はいつでもモデルを棄却することになるだろう．しかし，妥当な解析をするにはそのモデルで『十分（good enough)』なのだろう．仮にモデルが当てはまっていないのに，そのモデルに基づく推論をするのは妥当

訳注：適合度検定のこと．

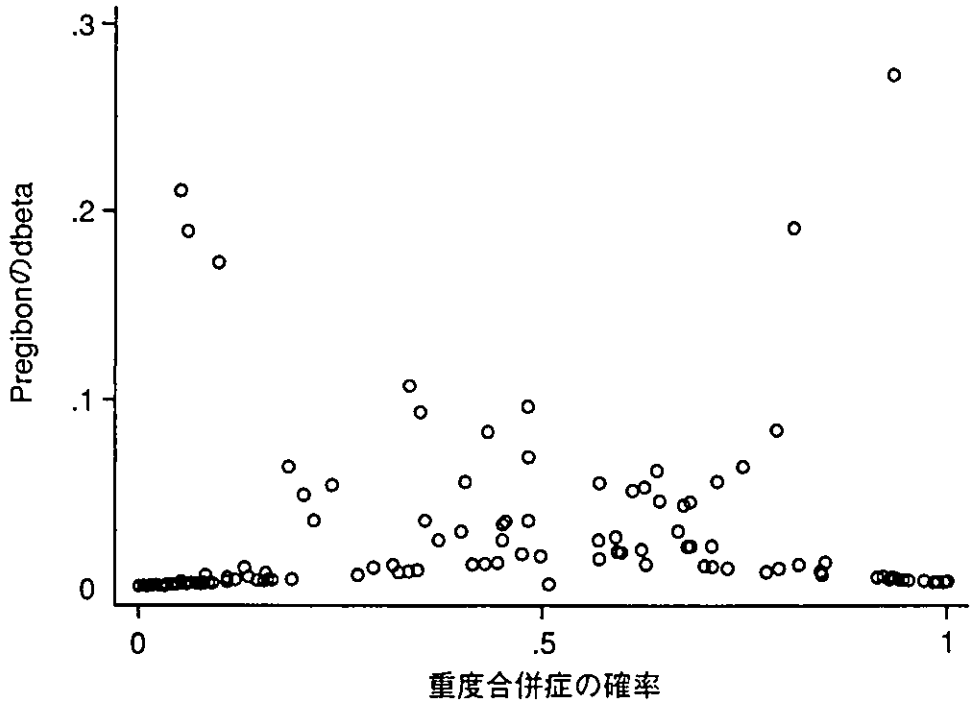

図 3.1　腹部手術データでのイベント推定確率に対する影響統計量のプロット

(C.Johnston から個人的に得た)

であるか？　一般的に言うと，答は『はい (yes)』である．しかし注意が必要なのである．

モデルをさらに確認するために，影響データを眺めることがある．これも多くのソフトウェアで利用できる．STATA では**総合的影響統計量** (overall influential statistic) があり，Pregibon の『dbeta』と書かれているのが使える．しかし，重回帰のように回帰係数の一つ一つについての影響統計量ではない．この dbeta を重度合併症の確率に対してプロットすると**図 3.1** のようになり，モデルの係数に影響しているデータが約 5 つあることが分かる．こうしたことは，もっと詳しく探求することができる．

ロジット・ロジスティック・対数線形

コンピュータマニュアルを読んでいると，**対数線形モデル** (log–linear model) という変わったモデルに出くわすことがあるかもしれない．こ

の対数線形モデルは大きな分割表の解析に用いる．ロジスティック回帰の代わりに，二値データの解析に使うことができる．少し昔のプログラムでは対数線形モデルが使えなかった．しかし，一般的にはこちらのほうがロジスティック回帰モデルより解釈するのが難しい．ロジスティックモデルとの違いは次の三つである．

- 従属変数と独立変数をあまり区別しない．
- ロジスティック回帰では独立変数に連続変数を含んでもよい．
- 対数線形モデルでは，従属も独立も全変数をモデルに含めないといけない．従属変数と独立変数との関連性は，交互作用項でもって見ることができる．したがって，対数線形モデルでは，Lavie[4] の研究例でやっているように，年齢を『若い (young)』と『年老いた (old)』(年齢は連続なので) に分けなければならない．それから高血圧の人の割合，高血圧でない人の割合に関するパラメータ，老人と若者の割合に関するパラメータを当てはめる．そのあとでこれら両者，つまり年齢と高血圧とが関係しているかを見る交互作用に関するパラメータを当てはめる．一方ロジスティック回帰では，高血圧の有無が文句なく従属変数になり，年齢が独立変数になる．

3.7 ケースコントロール調査

ケースコントロール調査の解析の中心はロジスティック回帰である．Swinscow 本[1] によると，オッズ比は X と Y を換えても変わらないという特徴を見た．すなわち，印刷屋の妻のほうが農家の妻よりも 3 ヵ月以上母乳を与えるかを見ることと，3 ヵ月以上母乳を与える人は農家の妻より印刷屋の妻のほうに多いこととは，オッズ比でいうと同じである．このように，因果を逆にする必要性がケースコントロール調査では生ずる．すなわち，疾病をもったケースを選択し，そして疾病をもっていないコントロールを選ぶのである．そこで疑っている原因に曝露した量に

ついて調べる．こうした考え方は，コホート研究とは異なっている．コホート研究では疑っている原因に曝露している人とそうでない人を捕らえ，疾病が起こるまで追跡するからである．

ロジスティック回帰を用いるとき，従属変数はケースなら1，コントロールなら0とコード化し，曝露に関する係数の推定値が対数オッズ比になる．もし疾病が稀なものなら，オッズ比は相対リスクの妥当な推定値となるだろう．

3.8 コンピュータ出力の解釈：マッチングを伴わないケースコントロール調査

Wald ら（1986）[7] を引用した Altman ら[6] で述べられた，4つのケースコントロール調査のメタアナリシスを考えよう．

表 3.5 4つの研究における女性の肺がん症例と対照ごとの受動喫煙への曝露状況[7]

	肺がん症例		対照		
研究	曝露	非曝露	曝露	非曝露	オッズ比
1	14	8	61	72	2.07
2	33	8	164	32	0.80
3	13	11	15	10	0.79
4	91	43	254	148	1.23

コンピュータ用に書き換えると

Y(症例)	n(症例＋対照)	曝露	研究
14	75	1	1
8	80	0	1
33	197	1	2
etc.			

上の表には8つの行がある．それは研究数×曝露の組み合わせになる．モデル上の従属変数はケース数である．さらに，各行においてケースとコントロールの総数も指定しなければならない．ロジスティック回帰プ

表 3.6　表 3.3 のケースコントロール研究に対するロジスティック回帰分析の結果

Logit estimates　　Number of obs = 977
LR chi 2(4) = 30.15
Prob > chi 2 = 0.0000
Log likelihood=−507.27463　　Pseudo R2 = 0.0289

_outcome	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lstudy_2	.1735811	.292785	0.593	0.553	−.4002669	.7474292
lstudy_3	1.74551	.3673518	4.752	0.000	1.025514	2.465506
lstudy_4	.6729274	.252246	2.668	0.008	.1785343	1.16732
exposed	.1802584	.1703595	1.058	0.290	−.1536401	.5141569
_cons	−1.889435	.2464887	−7.665	0.000	−2.372544	−1.406326

_outcome	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lstudy_2	1.189557	.3482845	0.593	0.553	.6701412	2.111565
lstudy_3	5.728823	2.104494	4.752	0.000	2.788528	11.76944
lstudy_4	1.959967	.4943937	2.668	0.008	1.195464	3.213371
exposed	1.197527	.2040101	1.058	0.290	.8575806	1.672228

ログラムの出力は**表 3.6** のようになる．ここで**研究**（study）とあるのは，4 水準のカテゴリー変数である．これは交絡因子であり，第 2 章で述べたように，3 つのダミー変数でモデル化できる．このことを**固定効果**（fixed effects）解析という．第 5 章にはこのような場合のダミー変数の使い方について，さらに詳しく述べている．プログラムには，対数オッズ（つまり回帰係数）またはオッズ比を選ぶオプションが付いている．肺がんと受動喫煙はオッズ比 1.198 で，95% 信頼区間 0.858〜1.672 で関連しているというのが主たる結果である．STATA で自動的に出てくる擬似 R2 は解釈するのが難しいので，引用すべきではない．出力は自動的にされるので，出力を読むときの注意点には違いない．

3.9　マッチングを伴うケースコントロール調査

マッチングを伴うケースコントロール調査では，各ケースは直接的に

一つ以上のコントロールとマッチングされる．正しく解析するには，このマッチングを考慮する必要がある．すぐ思いつく方法としては，マッチングされたグループに対して，それぞれ層とするダミー変数を当てはめるものである．しかし，これでは偏った推定値になることが示された[8]．そうではない方法が，**条件付きロジスティック回帰**（conditional logistic regression）というものである．単純な2×2表でいえば，これは**マクネマー検定**[1]（McNemar test）に等しくなる．最近のソフトウェアでは，ケースに対してコントロールが何例かあっても稼動するものがほとんどである．すなわち，ちょうど1：1マッチングである必要はないのである．

条件付き（conditional）尤度の考え方はかなり複雑であるが，簡単に述べてみよう．ケースに対してちょうど1例コントロールをマッチングされたケースコントロール調査で，(3.1) 式のようなロジスティックモデルを考えよう．ペア i に対して，ケースがイベント起こす確率を π_{i0}，コントロールがイベントを起こす確率を π_{i1} と仮定する．ペアの一方は**ケースでなければならない**（must）ことを，私たちは知っている．すなわち，ペアというのが**条件付き**（conditional）となり，ペアのうち一方しかイベント有りでないといけない．このとき，ケースがイベントを起こす確率は $\pi_{i0}/(\pi_{i0}+\pi_{i1})$ になる．例えば，ある夫婦チームが宝くじを当てたとしよう．夫は5枚券を買い，妻は1枚買ったことを知っていた．このとき，夫が宝くじに当たる確率はと尋ねられたら，妻よりも5倍高いだろうと言う．つまり，条件付き確率が1/6に対して5/6なのである．このようにして，ケース・コントロールのペアごとに，このような確率を掛け合わせて条件付き尤度が得られる．通常のロジスティック回帰と同様にして，この尤度を最大化するが，これは多くのソフトウェアで簡単に得られる．

モデルとしては (3.1) 式と同じだが，パラメータの推定法が異なる．ここでは条件無し尤度ではなく，条件付き尤度を使う．Swinscow本[1]で

さらに述べてあるが,ケースコントロールで同じという因子,例えばマッチング因子はモデル中の独立変数には出てこない.

3.10　出力の解釈:マッチングを伴うケースコントロール調査

データは Eason ら[9]から取り，Altman ら[6]で述べられている．喘息のため病院で亡くなった 35 名の患者に対して，それぞれ性と年齢でマッチした 35 名のコントロールが選ばれた．コントロールは，前の年に同じ病院を退院した者であった．患者モニタリングの充足度は独立に評価され，その結果は**表 3.7** に示された．

解析のため，35×2=70 行のデータにまとめられた．**表 3.8** に示したように，ケースやコントロール 1 例につき 1 行を割いている．最初のブロックはモニタリング充足度が不十分であった，10 名の死亡例と 10 名の生存例を表している．

条件付きロジスティック回帰の論理は,マクネマー検定と同じである．もしモニタリング充足度がケースおよびコントロールに同様ならば，そのペアはオッズ比には寄与しない．それが異なった結果のときにオッズ比に関係する．

表 3.7 から不十分なモニタリングのため病院で死亡する推定オッズ比は，二つの**不一致**（discordant）ペアの人数の比で与えられる．すなわち 13/3=4.33 となる．

表 3.7　喘息の死亡例 35 名とマッチングした生存例における病院でのモニタリング充足度[9]

		死亡例	
		非充足	充足
生存例（対照）	非充足	10	3
	充足	13	9

表 3.8　条件付きロジスティック回帰用に書き換えた表 3.7 のデータ

ペア番号	症例または対照(1=死亡)	モニタリング(1=不十分)
1	1	1
1	0	1
2	1	1
2	0	1
(10 ペアについて)		
11	1	1
11	0	0
12	1	1
12	0	0
(13 ペアについて)		
24	1	0
24	0	1
(3 ペアについて)		
28	1	0
28	0	0
(9 ペアについて)		

表 3.9　表 3.8 のマッチングを伴うケースコントロール研究での条件付きロジスティック回帰の結果

Conditional (fixed-effects) logistic regression　　Number of obs = 70

LR chi 2(1) = 6.74
Prob > chi 2 = 0.0094
Log likelihood = −20.891037　　Pseudo R2 = 0.1389

deaths	Odds Ratio	Std. Err.	z	P>| z |	[95% Conf. Interval]	
monitor	4.333333	2.775524	2.289	0.022	1.234874	15.20623

条件付きロジスティック回帰の結果は表 3.9 に与えられる.

ワルト検定の P 値は 0.022 なので，有意の結果である．このことはモニタリングが不十分だと，死亡リスクが高くなることを示唆している．尤度比の P 値は 0.0094 である．尤度比検定とワルト検定が異なっていることに注意しよう．その理由としては，表の中の数字が小さく，分布が離散なので，近似を用いた方法はすべて不正確になるためである．マ

クネマー・χ^2検定（スコア検定）の結果は $(13-3)^2/(13+3)=6.25$ だから，P=0.012 である．これは，尤度比検定とワルト検定の中間に位置する．それぞれの値は正しいものであるが，このように食い違う場合には検定手法を書く必要がある．多分一つ以上引用しておくのがよいだろう．第2章の直線回帰ではそうではなく，三つの方法がぴったり一致していた．

オッズ比は4.33で，95%信頼区間は1.23～15.21である．これはAltmanら[6]の66頁とは少し違っているが，そこでは小さな数字に適している**正確法**（exact method）を用いたためである．

単純なマクネマー検定よりも条件付きロジスティック回帰のほうがすぐれるのは，他の共変量をモデルに取り込めるからであろう．上の例でいうと，70人すべてに**気管支拡張器**（bronchodilators）使用の有無を取ってあれば，それを病院で死亡するリスクファクターとして組み込むことができる．

3.11　条件付きロジスティック回帰の実際

Churchillら[10]は，マッチングを伴うケースコントロール調査を実施した．ケースは過去3年間に妊娠したことのある10代の女性であった．ケースの年齢に最も近くて，10代には妊娠をしなかった女性3名をコントロールとした．このコントロールも同じ診療所で見つけられた．条件付きロジスティック回帰で解析した結果，ケースのほうがコントロールよりも妊娠の前年に相談に来ることが多かった（オッズ比2.70，95%信頼区間：1.56～4.66）．

3.12　ロジスティック回帰の結果を報告するには

●ロジスティック回帰の結果には，データ数，説明変数の係数とSE，あ