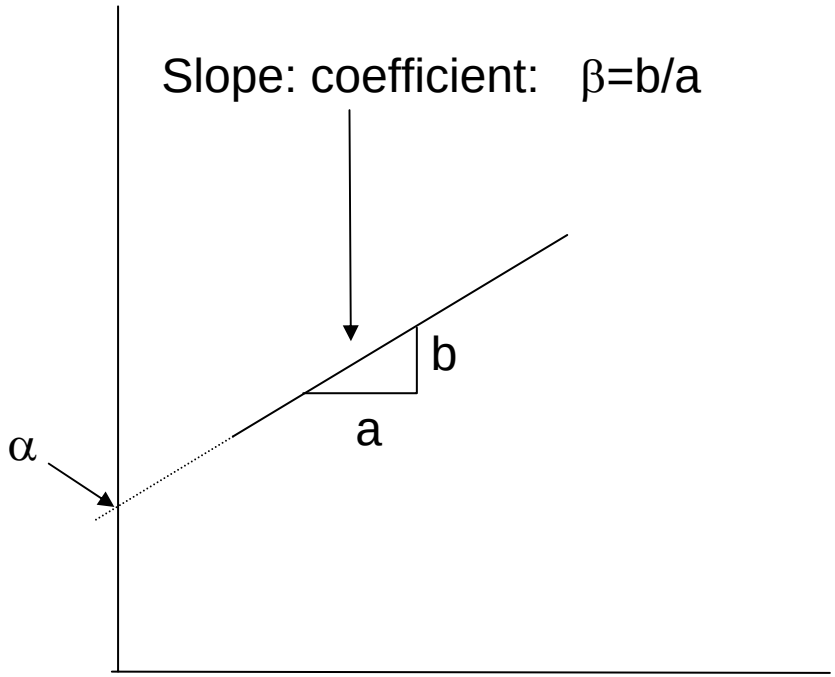
回帰

Simple Linear Regression

- 相関関係(correlation)のように2つの連続して変化するx とy という変数を比較する際、simple linear regression が用いられる。相関関係の解析においては x と y を入れ替えても問題ないが、simmple linear regression model ではx がexplanatory variable として変化すると、yもresponse として変化する。よってxをindependent variable, yをdependent variable とも表現する。そして1つの直線を引くことにより新たに与えられた値 x からy を算出 (予測) することができる (図1) 。
- この
- m y/x = a + bx
- を使って表される公式は不変の真理である。極端な話、世界全てのデータを集めたものをmとする。しかしこれを知ることは実際不可能である。そこで我々は手元にあるデータからこの世界のデータを推論するわけである。しかしデータ数が十分でなければ、世界のデータから少しずれてしまうかもしれない。その分をe (error)として表現する。yは1部のサンプルを用いていることを示している。
- Simple linear regression において、変数Xは観察された範囲において変数Yと直線的関係にあることが期待される。そうでなければsimple linear regression にはならない。またXとYはそれぞれ独立しながらも同じ分布を示す。そうでなければやはりsimple linear regression にならない。これは仮定である。もちろん実際は直線に近くなることはあっても直線にはならないことがほとんどである。



$\mu_{y/x} = \alpha + \beta x$

$y = \alpha + \beta x + \varepsilon$

図1. Simple linear regression **における**β **と**α **、そして**ε**の意味するもの**

The Method of Least Squares

- 成人男性における肥満度と収縮期血圧の関係をみてみよう（図2）。肥満度が増加すると血圧も上昇する傾向にある。これらの点をよく表すように線を引くにはどうしらたよいだろうか？2人の人に直線を引かせる微妙に異なることだろう。そこで皆の引く線が一致すように我々はmethod of least squares という概念を用いる。

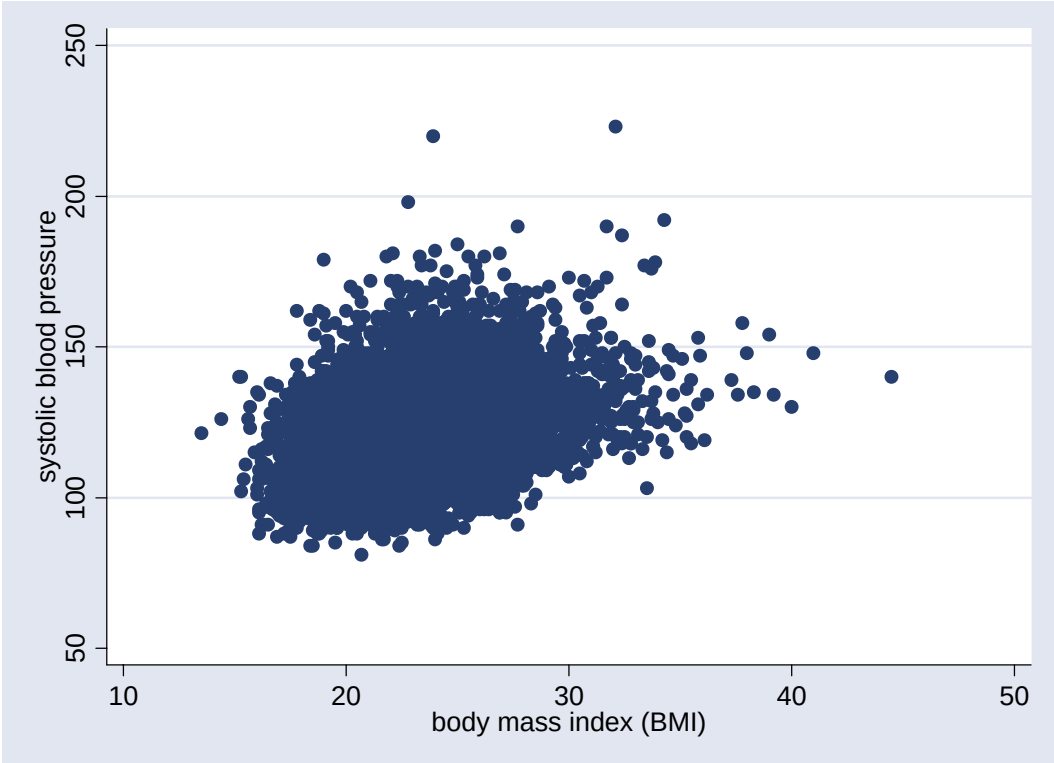
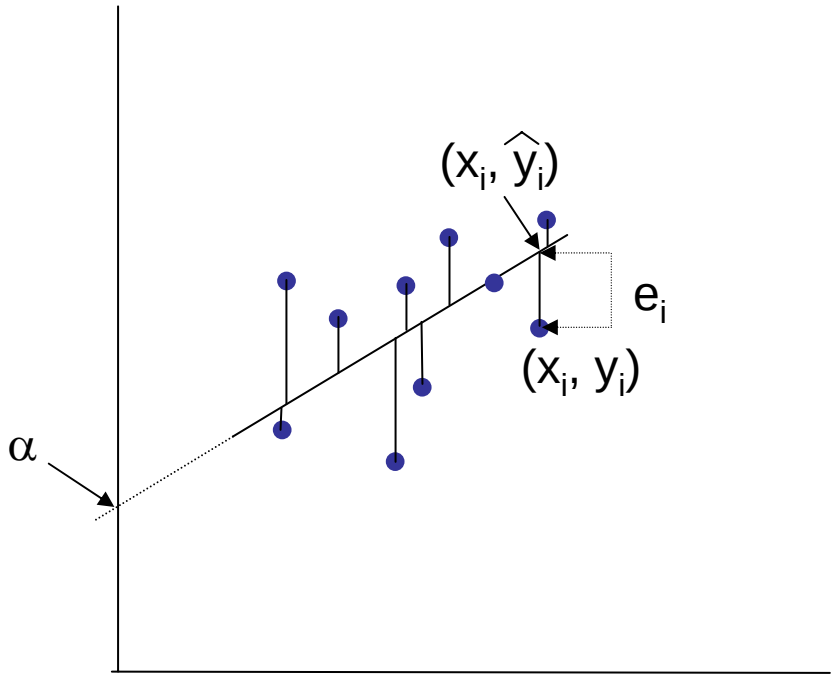

systolic blood pressure
250
200
150
100
50
10
20
30
40
50
body mass index (BMI)

図2　肥満と血圧の関係

- グラフ上のどの点(xi, yi)も描こうとしている直線からは一定の距離(ei)をもって存在している（中には直線上に乗るものもあるかもしれない）。そこでei の2乗の総和が最小になるように直線を引く（プラスとマイナスのものが存在するため、二乗してやる）（図3）。



$\mu_{y/x} = \alpha + \beta x$

$y = \alpha + \beta x + \varepsilon$

図3. Simple linear regression **における**β **と**α **、そして**ε**の意味するもの**

- しかしこれを計算するのは至難の技で、コンピュータにやってもらう。そうすると、傾き(b)とY軸と交差するYの値(a)が算出される（図4）。
- b: 1.54, a: 85.3
- 収縮期血圧＝1.54 x BMI + 85.3
- t値はcoefficiency をstandard error で割ったものである。またR-squared はrの二乗、すなわち相関係数の二乗なので、ここで相関係数は$\sqrt{}$0.1 = 0.3となる。相関係数(Pearson’s correlation coefficient)が –1 から＋1までの範囲であるのに対して、R2は0 から1までである。ここではR2 は0.1であるが、文章にすると「出生時平均血圧の10%は妊娠週数との直線関係で説明される」ことを意味している。

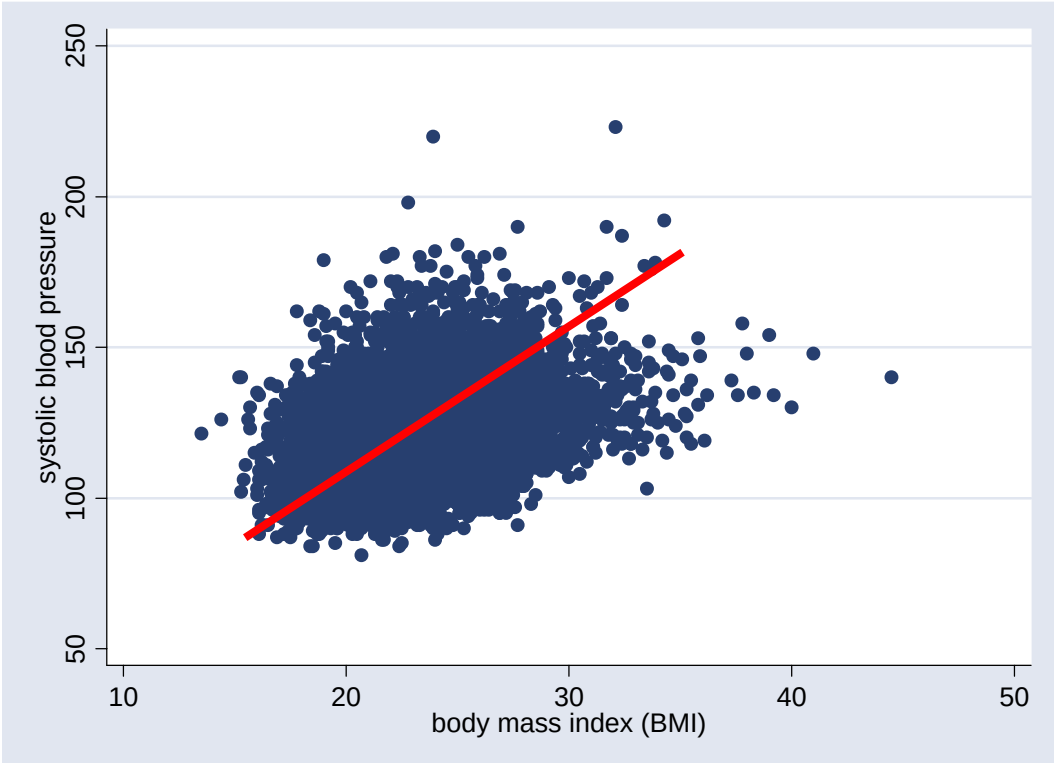


図4　肥満と血圧の関係: least square method により得た直線を示す。

Multiple Regression

- simple linear regression **では2つの変数を比較したが、もっと多くの変数を採りいれたいと思う。3つ以上の変数を入れた解析方法が**multiple regression **である。3つの変数であれば**3**次元、4つであれば4次元となり、もはやグラフに示すことはできなくなってしまう。しかし、いくつかの変数を入れてある結果を予測しうるのでいろいろなものに使うことができる。**
-
- $y = a + b1x1 + b2x2 + ... + bqxq + e$
-
- Random error associated with y
-
- **原理は**simple linear regression **と同じで**least squares **を用いる。先の例で、**BMI**と**LDL**コレステロール**(LDL)、HDL**コレステロール**(HDL)、中性脂肪**(TG)、**ヘモグロビン**A1c(HbA1c)**が収縮期血圧**(sbp)**のどのような影響を与えているかを検討してみた。**

- **コンピュータ解析により以下の結果を得たとしよう。**sbp＝1.61 x BMI + 0.013 x LDL + 0.18 x HDL + 0.019 x TG + 1.69 x HbA1c + 60.5
- **仮にAさんの健康診断の結果が以下の通りだとする。**
- BMI LDL HDL TG HbA1c
- 20.3 96 45 135 4.7
- sbp＝1.61 x (20.3) + 0.013 x (96) + 0.18 x (45) + 0.019 x (135) + 1.69 x (4.7) + 60.5
- = 113.1 mmHg
- **Aさんの血圧は**113 mmHg **と予想される。**
- **結果は血圧のように連続変数を用いることが多いが、変数は連続変数でもカテゴリーでも0や1で表されるようなものでもかまわない。また、個々の変数をかけあわせて相乗効果をみる、連続変数を適当な閾値で切る、あるいはカテゴリーで分けることも有用かもしれない（線形モデル** vs. **閾値モデルを参照）。そして、**R2**が大きくなればなるほど、解析モデルの変数は結果をよく表していることになる。**

モデルの評価

- coefficient biが0と違わない（有意差がなくH0を棄却できない）ときにはその変数を捨ててモデルを簡単にすることができる。変数が非常に多くある場合我々はどれを加えてどれを捨てるか決めなくてはならないわけであるが、これは統計的および非統計的判断により決定する。サンプル数に比べてあまりにも変数が多いとb値の動揺が激しくなる。最初に我々はどのような変数を検討するか、今までの知識を駆使して検討する。例えば血圧に影響を与える因子を考えた時、何が影響しそうかのアイデイアが無いとデータを収集できない。
- 解析時にできることは可能性のある変数を全ての組み合わせを用いて検討する。これをall possible models と呼ぶ。変数が少ないときは何とかなるかもしれないが、変数が多いときは先に述べたようにb値が不安定になる。そこで我々は2種類のstepwise approach を用いる。1つはforward selection で、我々は最も大きなcoefficient をもつ変数からはじめ、次々に変数を加え、そしてR2の変化を認めなくなった時点で終了する（統計学的にモデルを評価し、カイ二乗検定で比較する）。一方backward eliminationでは、最初に有意な変数を全て加えておいて1つずつ減らしていく。

- Forward selection の逆であるから、R2がほとんど変化しない場合その変数を抜くことができる。多分に感覚的な要素が入るため、解析を行なった人、Forward or backward によって最終的な数式が異なる可能性があり得る。統計ソフトによっては、コマンド入力によりこのプロセスを自動的に行ってくれる。
- 多少式が異なっても大きな問題にはならないことが多いのであるが、我々が注意しなくてはならないのはcollinearity である。Collinearity は2つ以上の変数がお互い強く連動するとき発生する。例えば喘息発作の発生と環境要素を検討しようとする際、気圧配置とオゾン濃度とその日の運動量であれば、ほぼ独立した変数として解析できると思われるが、車の交通量の多寡は大気汚染の程度に直接影響してくるから、collinearity を生じる可能性がある。Collinearity を生じるとcoefficient and/or standard error の変動が大きくなる。例として、表1を示す。
- この表ではx2を加えることによってR2は変化していないのにcoefficient は倍に、standard error は10倍になっている。このように特にstandard error が大きく変動するときにはcollinearity を考慮しなくてはならない。このような場合x1とx2の間のcorrelation を調べてみればすぐにわかることである。そしてxiはx2を加えることによってpが0.05を超えて有意性を失ってしまう。よってある変数が他の変数との関係でcollinearity にあたると判断したら、その変数を除外する。

モデルの評価

表1　collinearity の性癖

	Variable x_1	Variable x_1 + x_2
Coefficient	-1.412	-2.815
Standard error	0.406	4.985
Test statistic	-3.477	-0.565
p-value	0.001	0.574
R^2	0.653	0.653
Adjusted R^2	0.646	0.642

Logistic Regression model

- 臨床において我々はしばしばyes/no で回答を求められる。その最たるものが生死であろうし、患者さんにとって治るか治らないかが最も知りたいところである。このようなyes/no であらわされる変数をdichotomous とよび、コンピュータでは1/0 として表現される。そしてdichotomous y を扱う場合はlinear regression でなく、logistic regression を用いて解析する（結果が0/1 だけではなく、0/1/2/3/4/5などであってもlogistic regression で解析することは可能であるが、ここでは述べない）。
- 例えば2500g未満の低出生体重児が生まれるか否かを諸々の因子から予測したいと思う。「母親の最終月経時（妊娠直前）の体重が軽い程低出生体重児を出産しやすい」という仮説を検証してみる。
- 低出生体重児は189例中59例であり、コンピュータ上の生データとしては1で表現する。一方、正常体重児は0で示す。その比率は0.312である。もしも我々の「母親の体重が軽ければ低出生体重児になりやすい」という仮説が正しいならば、これを公式として表して、次に妊婦の体重を調べて低出生体重児を出産す確率を具体的数値をもって妊婦に提示する事ができる。まずは母親の体重をx軸に、低出生体重児であるか否かをy軸に示す（図5）。

Logistic Regression model

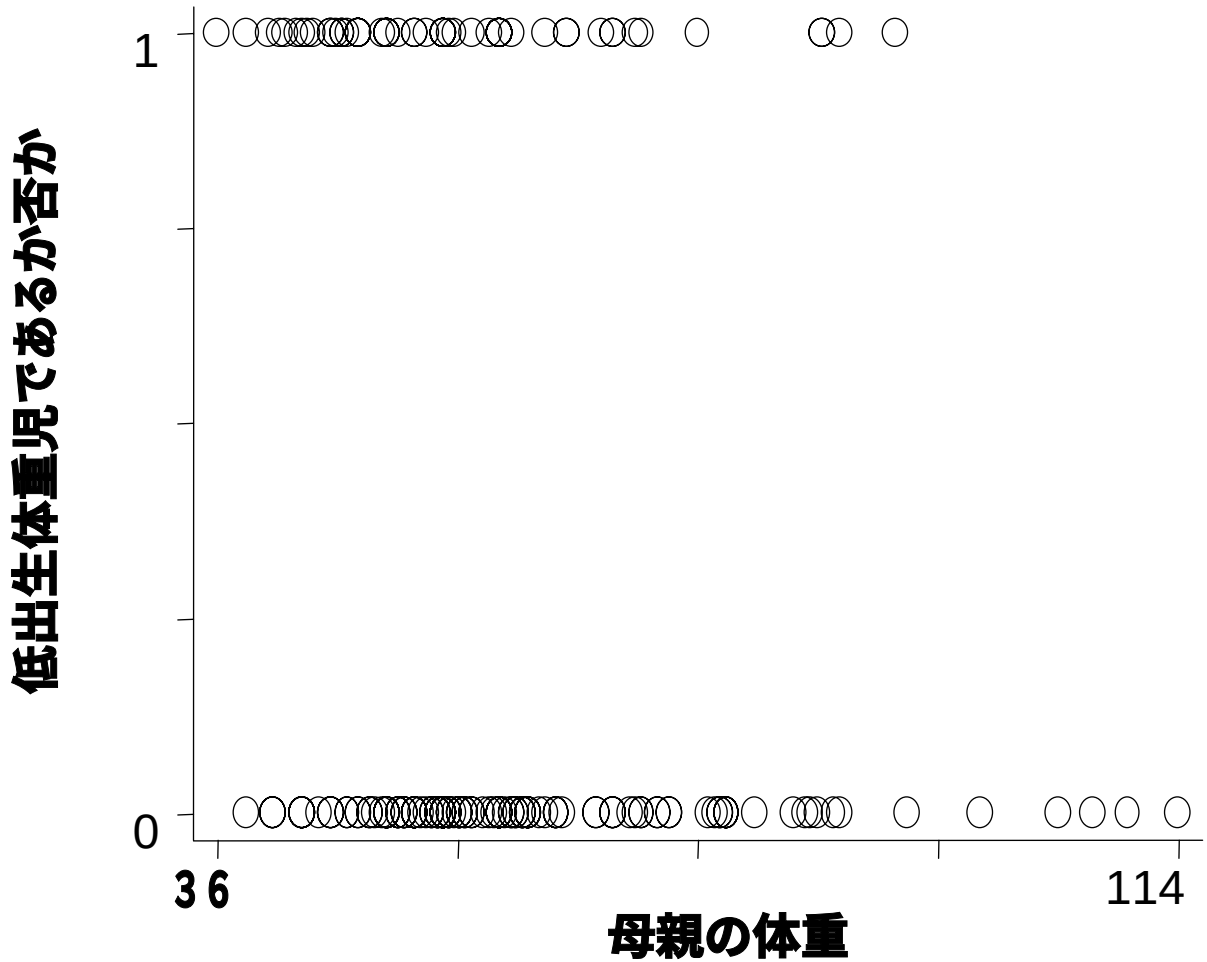


図5　母親体重と「低出生体重児であるか否か」の関係

- **yが0か1であるため、図のような変わったグラフとなってしまった。若干体重が重い方が正常体重児を出産しやすいようにもみえる。ここで低出生体重児が生まれる確率をpとする。まずはsimple linear regression model にあてはめてみるとどうか？**
-
- p = a + b1x1
-
- **とすると、確率pは0‐1の間の値をとるべきなのに、上の公式では1を超えてしまったり、マイナスの値になったりしてしまう。**
-
- $p = e^{a + b1x1}$
-
- **としたらマイナスにはならないが、やはり1を超えてしまう。**

- $p = e^{a + b1x1} / (1 + e^{a + b1x1})$
-
- こうすればpは0から1の範囲に収まる。これをlogistic function と呼ぶことにする。この公式は以下のように変換することもできる。
-
- $P/ (1 – p) = (e^{a + b1x1} / (1 + e^{a + b1x1})) / [1/ (1 + e^{a + b1x1}) = e^{a + b1x1}$
- $\text{Ln} [p/ (1 – p)] = a + b1x1$
-
- 式の右はlinear regression のものと同じである。左はodd のlog である。つまり低出生体重児が生まれる確率pのoddの自然対数は直線で表すことができるのである。これをlogistic regression と呼ぶ。Linear regression と違ってlogistic regression ではleast squares の原理を適応することができない。そのかわりmaximum likelihood estimation を用いる。

- **コンピュータを用いて母親の体重と低出生体重児出産の確率との関係は以下のようになる（図6）。**
-
- ln[p/(1 – p)] = 0.998 – 0.031x(**母親体重**)
-
- **例えば母親の体重が**45kg**だったとする。**
-
- ln[p/(1 – p)] = 0.998 – 0.031x45 = -0.397
- p/(1 – p) = e-0.397=0.672
- p = 0.672/ (1+0.672) = 0.40
-
- **よって**45kg**の母親が低出生体重児を生む確率は**40**%になる。**

Logit estimates　　Number of obs = 189
LR chi2(1) = 5.98
Prob > chi2 = 0.0145
Log likelihood = -114.34533　　Pseudo R2 = 0.0255

低出生体重児	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
母親の体重	-.0309653	.0135894	-2.279	0.023	-.0576	-.0043307
_cons	.9983142	.7852888	1.271	0.204	-.5408236	2.537452

図6　母親体重と「低出生体重児であるか否か」の関係をコンピュータ(STATA 7.0)でロジスティック解析したもの

- 母親の体重をx軸に、低出生体重児を出産す確率pをy軸にとったグラフである（図7）。その関係は直線ではない。

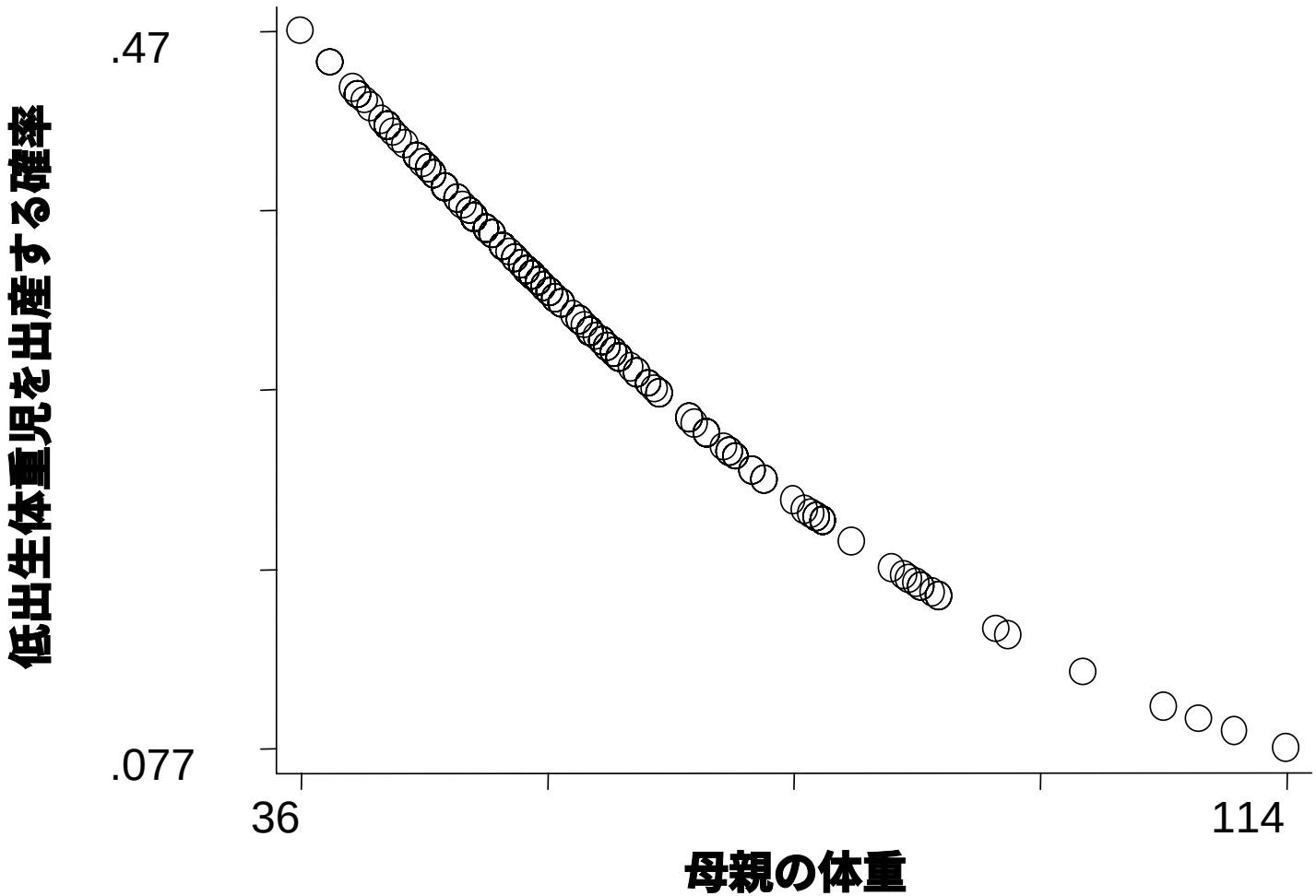


図7　母親体重と「低出生体重児を出産する確率」の関係

Multiple Logistic Regressions

- 今までは母親の体重という変数を1つしか考えなかった。しかし実際の臨床の場おいて結果を左右する因子は山ほどあるのが普通である。低出生体重児を出産すリスクファクターとして妊娠中毒症は子宮内発育不全を来たしやすいため重要である。もしも低出生体重児出産の確率を計算するのに、いくつもの予後因子を同時に解析できると大変便利である。これがmultiple logistic regression である。simple linear regression と multiple regression の関係と同じことである。
-
- ln[p/(1 – p)] = a + b1x1 + b2x2 + ... + bnxn
-
- 用いる変数は必ずしも連続変数である必要はない。例えば母親の喫煙(smoke: 0 or 1)、妊娠中の労働状態(plt: 0 or 1)、妊娠中高血圧(ht: 0 or 1)、子宮被刺激性(uterine irritability: ui: 0 or 1)、妊娠初期3ヶ月における産科医受診回数(ftv)なども変数として扱える。
- ここではコンピュータで解析結果を示す（図8）。

```
. logit low age we smoke ptl ht ui ftv
Logit estimates                           Number of obs   =        189
                                          LR chi2(7)      =      25.92
                                          Prob > chi2     =     0.0005
Log likelihood =  -104.3764               Pseudo R2       =     0.1104
------------------------------------------------------------------------------
         low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0432489   .0354043    -1.222   0.222      -.1126399    .0261422
          we |  -.0316464   .0146579    -2.159   0.031      -.0603753   -.0029174
       smoke |   .5539317    .344437     1.608   0.108      -.1211525    1.229016
         ptl |   .5943356   .3482606     1.707   0.088      -.0882425    1.276914
          ht |    1.87316   .6908402     2.711   0.007       .5191376    3.227182
          ui |   .7393009   .4566633     1.619   0.105      -.1557427    1.634344
         ftv |   .0234335   .1731271     0.135   0.892      -.3158894    .3627564
       _cons |   1.390719    1.09008     1.276   0.202      -.7457992    3.527238
------------------------------------------------------------------------------
```

図8　7つの変数をmultiple logistic regression analysis した際の結果　STATA 7.0

- **母親の体重と高血圧が低出生体重児出産と統計学的有意性をもって関係していることがわかった。それではこの2つだけで解析してみる（図9上）。**
- **もしも以下の公式を得、母親の体重が**45kg**で高血圧がなければ**
-
- ln(p/(1 – p)) = 1.45 + 1.86*(0) – 0.041*45 = -0.395
- p/(1 – p) = 0.67, p = 0.67/(1+0.67) = 0.40
-
- **先の結果と同じである。**
- **もしも母親の体重が**45kg**で高血圧があれば**
-
- ln(p/(1 – p)) = 1.45 + 1.86*(1) – 0.041*45 = 1.465
- p/(1 – p) = 4.33, p = 4.33/(1+4.33) = 0.812
-
- **同じ体重でも高血圧を合併すると低出生体重児を出産す確率が非常に高くなる。**
-

- 今までみてきたcoefficient はOR = expcoefficient で転換できる。例えばexp1.86 = 6.4,
- つまり体重が同じ人で高血圧のある人とない人を比べると、低出生体重児を出生すリスクは6.42倍違うといえる（図9下）。
- 最初は7つの変数を用いてmultiple logistic regression 解析を行った。そして、統計学的に有意であった2つの因子をとりだして再度multiple logistic regression 解析を行った。どちらの解析を示せばよいのだろうか？7つの変数を用いた場合、Log likelihood = -104.3764 であり、2つの変数をもちいた場合のそれは、Log likelihood = -110.57105 であり、より0に近い方を選ぶ。
-
- 2 x [log-likelihood(null) – log likelihood(alt)]
-
- これをc2で検定することもできる。自由度はそれぞれの表の変数の数の差になる。全ての変数を含むモデル(alt)の方が、有意な2つの変数を含むモデル(null)より優れていた(p = 0.03)。このようにモデル全体として最も効率的に結果発生を予測できるモデルを選択することもできるが、表に掲載する変数のp 値を一定のところでカットオフする方法もある。そのカットオフ値に定説はないが、「p値が0.2未満であればその変数を加えるべき」という意見がある。
-

```
Logit estimates                           Number of obs   =        189
                                          LR chi2(2)      =      13.53
                                          Prob > chi2     =     0.0012
Log likelihood = -110.57105               Pseudo R2       =     0.0577
------------------------------------------------------------------------------
         low |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ht |   1.855511   .7009715      2.647   0.008       .4816325     3.22939
          we |  -.0410851   .0145243     -2.829   0.005      -.0695523   -.0126179
       _cons |   1.450679   .8209564      1.767   0.077      -.1583656    3.059724
------------------------------------------------------------------------------
Logit estimates                           Number of obs   =        189
                                          LR chi2(2)      =      13.53
                                          Prob > chi2     =     0.0012
Log likelihood = -110.57105               Pseudo R2       =     0.0577
------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ht |   6.394967    4.48269      2.647   0.008       1.618715    25.26425
          we |   .9597474   .0139397     -2.829   0.005       .9328113    .9874614
------------------------------------------------------------------------------
```

図9　7つの変数から有意であった、2つの因子（母親高血圧、母親体重）だけをmultiple logistic regression で調べた。Aはcoef (= coefficiency) であり、下はOdds Ratio になっている点に注目。ただ、Ln(OR) = coefficiency の関係であるので、同じ結果をみているに過ぎない。