Logistic Regression model

Linear regression model において response Y は以下の公式で示されます。

$y = \alpha + \beta_1 x_i + \beta_2 x_2 + \ldots\ldots + \beta_q x_q + \varepsilon$

linear regression では response y は連続変数でした。例えば妊娠週数を x としたときの出生体重 y は連続変数です。しかし臨床において我々はしばしば yes/no で回答を求められます。その最たるものが生死でしょうし、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す。このような yes/no であらわされる変数を dichotomous とよび,コンピュータを使用する際 1/0 として表現されます。そして dichotomous y を扱う場合は linear regression でなく、logistic regression を用いて解析します。

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います。「母親の最終月経時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ます。

```
. sum we

Variable |     Obs       Mean   Std. Dev.      Min       Max
---------+--------------------------------------------------
      we |     189    58.93593   13.88304     36.32     113.5
```

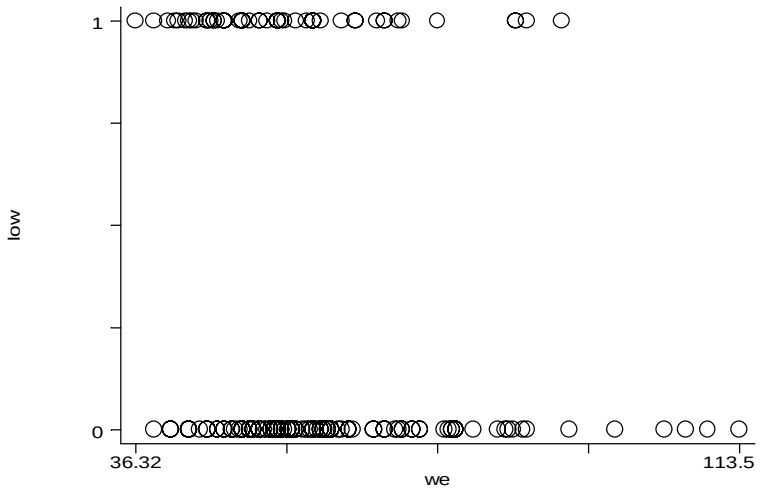
.母親の平均体重は 58.9±13.9kg です。

```
. tab low

       low |      Freq.     Percent       Cum.
------------+-----------------------------------
         0 |       130       68.78      68.78
         1 |        59       31.22     100.00
------------+-----------------------------------
     Total |       189     100.00
```

低出生体重児は 1、正常体重児は 0 で示されています。189 例中 59 例でした。その比率は 0.312 です。もしも我々の仮説が正しく母親の年齢が若ければ低出生体重児にな

りやすく、これを公式として表すことができれば、次に母親の年齢を聞いて低出生体重児を出産する確率を具体的数値をもって示す事ができます。まずは母親の年齢をｘ軸に、低出生体重児であるか否かをｙ軸に示します。



ｙが０か１であるため、上のような変わったグラフとなってしまいました。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ます。

ここで低出生体重児が生まれる確率をｐとします。まずは simple linear regression model にあてはめてみるとどうでしょうか？

$$p = \alpha + \beta_1 x_1$$

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います。

$$p = e^{\alpha + \beta 1 x 1}$$

としたらマイナスになりませんが、やはり１を超えてしまいます。

$$p = e^{\alpha + \beta 1 x 1}/(1 + e^{\alpha + \beta 1 x 1})$$

こうすればｐは０から１の範囲に収まります。これを logistic function と呼びます。この公式は以下のように変換することもできます。

$P/(1 - p) = (e^{\alpha + \beta 1x1}/(1 + e^{\alpha + \beta 1x1}))/[1/(1 + e^{\alpha + \beta 1x1}) = e^{\alpha + \beta 1x1}$

$Ln[p/(1 - p)] = \alpha + \beta_1 x_1$

式の右は linear regression のものと同じです。左は odd の log です。つまり低出生体重児が生まれる確率 p の odd の log は直線で表すことができるのです。これを logistic regression と呼びます。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ません。そのかわり maximum likelihood estimation を用います。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ります。

```
. logit low we

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -114.41626
Iteration 2:   log likelihood = -114.34546
Iteration 3:   log likelihood = -114.34533

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

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
          we |  -.0309653   .0135894    -2.279   0.023        -.0576   -.0043307
       _cons |   .9983142   .7852888     1.271   0.204     -.5408236    2.537452
------------------------------------------------------------------------------
```

$ln[p/(1 - p)] = 0.998 - 0.031x$(母親体重)

つまり母親の体重が 10kg 違うと、$ln[p/(1 - p)]$も 0.31 違うことになります。例えば母親の体重が 45kg だったとします。

$ln[p/(1 - p)] = 0.998 - 0.031x45 = -0.397$

$p/(1 - p) = e-0.397=0.672$

$p = 0.672/(1+0.672) = 0.40$

よって 45kg の母親が低出生体重児を生む確率は 40%になります。

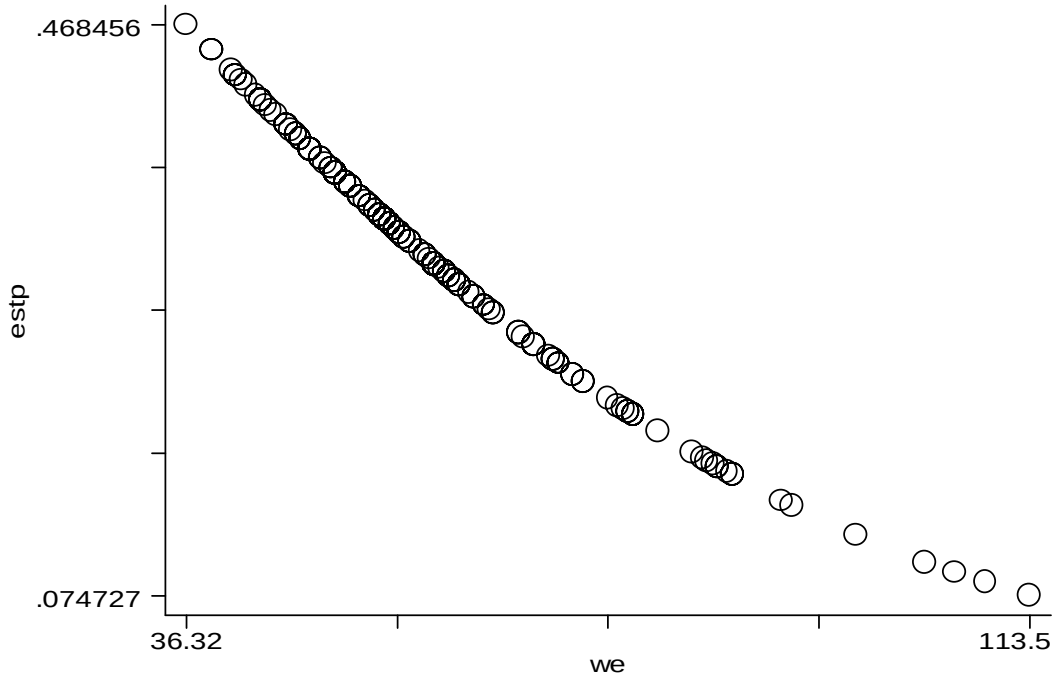
```
. gen lnp1p = .9983142  -.0309653*we

. gen p1p = exp(lnp1p)

. gen estp = p1p/(1+p1p)

. graph estp we

.
```



母親の体重を x 軸に、低出生体重児を出産する確率 p を y 軸にとったグラフです。その関係は直線ではありません。

Multiple Logistic Regression

今までは母親の体重という変数を１つしか考えませんでした。しかし実際の臨床のおいて結果を左右する因子は山ほどあるのが普通です。低出生体重児を出産する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す。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だと思いませんか？これが multiple logistic regression です。パターン的には simple linear regression と multiple regression の違いを連想していただければいいと思います。

$\ln[p/(1-p)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

logistic regression model において用いる変数は必ずしも連続変数である必要はありません。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ます。このような変数を explanatory variable と呼びます。

```
. 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
```

```
------------------------------------------------------------------------------
```

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りました。それではこの２つで解析してみます。

```
. logit low ht we

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.74314
Iteration 2:   log likelihood = -110.57165
Iteration 3:   log likelihood = -110.57105

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
------------------------------------------------------------------------------
```

ln(p/(1 – p)) = 1.45 + 1.86*(高血圧) – 0.41*（母親体重）

もしも母親の体重が 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

先ほどと同じ結果です。
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る確率が 80%を超えてしまいます。

今までみてきたcoefficientはOR＝exp^{coefficient}で転換できます。STATA logisticコマンドを用いれば簡単にORとその 95%CIを求めることができます。
例えばexp$^{1.86}$ = 6.42,
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るリスクは 6.42 倍違うといえます。

```
. logistic low ht we

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
------------------------------------------------------------------------------

. logistic low ht we age smoke ptl 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 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ht |   6.508829   4.496561     2.711   0.007     1.680578    25.20851
          we |   .9688492   .0142013    -2.159   0.031     .9414112    .9970868
```

```
    age |    .957673   .0339057    -1.222   0.222      .8934723    1.026487
  smoke |   1.740081   .5993484     1.608   0.108      .8858989    3.417864
    ptl |   1.811827   .6309878     1.707   0.088      .9155388    3.585557
     ui |   2.094471   .9564679     1.619   0.105      .8557794    5.126097
    ftv |    1.02371    .177232     0.135   0.892      .7291401    1.437286
------------------------------------------------------------------------------
```

上に２つの表を示しました。最初の表は統計学的に coefficient (or OR)が有意であった変数、高血圧と母親の体重、のみを変数として用いたものです。次のものは統計学的に有意でない covariate も含めてみました。Logistic regression の場合、log likelihood が最も小さい数値のものが、より予測される値を示すことになりますから、この場合には log likelihood の小さい全ての変数を含んだ最後のモデルを使用するべきです。もしも 2 つの表の log likelihood が近接していて差があるかどうかわからないときには chunk test or log-likelihood ratio test を行ないます。

```
2 x [log-likelihood(null) – log likelihood(alt)]
```

これをχ^2で検定します。自由度はそれぞれの表の変数の数の差になります。または最初の表の後にlrtest, saving(0) ２つめの表の後にlrtest, saving(alt) そしてlrtest, model(0) using(alt) とすると自動的上の計算を行なってｐ値を算出してくれます。

```
. logistic low ht we

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
------------------------------------------------------------------------------
```

```
. lrtest, saving(0)

. logistic low ht we age smoke ptl 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 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ht |   6.508829   4.496561     2.711   0.007     1.680578    25.20851
          we |   .9688492   .0142013    -2.159   0.031     .9414112    .9970868
         age |    .957673   .0339057    -1.222   0.222     .8934723    1.026487
       smoke |   1.740081   .5993484     1.608   0.108     .8858989    3.417864
         ptl |   1.811827   .6309878     1.707   0.088     .9155388    3.585557
          ui |   2.094471   .9564679     1.619   0.105     .8557794    5.126097
         ftv |    1.02371    .177232     0.135   0.892     .7291401    1.437286
------------------------------------------------------------------------------

. lrtest, saving(alt)

. lrtest, model(0) using(alt)
Logistic:  likelihood-ratio test                  chi2(5)     =     12.39
                                                  Prob > chi2 =    0.0298
```

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)より優れていることになります。変数をモデルに加えるか否かは一定の見解を得ていませんが、「p 値が 0.2 未満であればその変数を加えるべき」という意見があります。