Multiple Regression

simple linear regression では２つの変数を比較しましたが、もっと多くの変数を採りいれたいと思います。３つ以上の変数を入れた解析方法が multiple regression です。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います。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ます。

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots + \beta_q x_q + \varepsilon$$

random error associated with y

原理は simple linear regression と同じで least squares を用います。先の例で低出生体重児の収縮期血圧と出生 5 分の apgar score を比較してみましょう。

```
. regress  sbp apgar5
```

```
  Source |       SS       df       MS              Number of obs =     100
---------+------------------------------            F(  1,    98) =    2.22
   Model |  284.675502     1  284.675502            Prob > F      =  0.1398
Residual |  12588.6845    98  128.455964            R-squared     =  0.0221
---------+------------------------------            Adj R-squared =  0.0121
   Total |   12873.36    99  130.033939            Root MSE      =  11.334

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
   apgar5 |   .6977341   .4686968     1.489   0.140    -.2323795    1.627848
   _cons |   42.71916   3.140968    13.601   0.000     36.48601    48.95231
------------------------------------------------------------------------------
```

R-squared = 0.0221 で相関関係は弱いようです。２変数の間のRはsimple correlation を示していますが、３つ以上の変数の場合にはmultiple correlation といえます。それでは妊娠週数との関係をみてみましょう。Regress y x_1 x_2 x_3. . . . のようにタイプします。

```
. regress  sbp apgar5 gestage
```

```
  Source |       SS       df       MS              Number of obs =     100
```

```
---------+------------------------------           F(  2,    97) =    4.76
   Model |  1151.36376     2  575.681879           Prob > F     =  0.0106
Residual |  11721.9962    97  120.845322           R-squared    =  0.0894
---------+------------------------------           Adj R-squared =  0.0707
   Total |   12873.36    99  130.033939           Root MSE     =  10.993

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
  apgar5 |   .4875149   .4613278     1.057   0.293     -.4280931    1.403123
 gestage |   1.184826   .4424232     2.678   0.009      .3067381    2.062913
   _cons |   9.803418   12.66293     0.774   0.441     -15.32899    34.93583
------------------------------------------------------------------------------
```

収縮期血圧 = 9.803418 + 1.184826 x 妊娠週数 + .4875149 x apgar score
となります。妊娠が 31 週で apgar score 7 点のベビーの予測される収縮期血圧はどれくらいになりますか？9.803418 + 1.184826 x 31 + .4875149 x 7 = 45 mmHg です。
apgar score を加えたことによって多少 R-square と MSE は改善しました。その他の因子を加えてみましょう。妊娠中毒症のデータを加えてみましょう。妊娠中毒があれば１、なければ０です。このような変数を Dummy variables あるいは Indicator variables と呼びます。

```
. regress  sbp tox gestage apgar5

  Source |       SS       df       MS              Number of obs =     100
---------+------------------------------           F(  3,    96) =    3.24
   Model |   1183.5956     3  394.531868           Prob > F     =  0.0255
Residual |  11689.7644    96  121.768379           R-squared    =  0.0919
---------+------------------------------           Adj R-squared =  0.0636
   Total |   12873.36    99  130.033939           Root MSE     =  11.035

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
     tox |  -1.531149   2.976063    -0.514   0.608     -7.438586    4.376289
 gestage |   1.285027   .4849377     2.650   0.009      .3224333    2.247621
  apgar5 |   .4978851   .4635248     1.074   0.285     -.4222044    1.417975
   _cons |   7.165329   13.70645     0.523   0.602     -20.04177    34.37242
```

相関係数は 0.3 です。少し改善してきました。

さらに性別、grmhem を加えるともう少しだけ R-square の改善をみました。

```
. regress  sbp sex tox grmhem gestage apgar5

  Source |       SS       df       MS              Number of obs =     100
---------+------------------------------           F(  5,    94) =    2.34
   Model |  1425.08678     5  285.017356            Prob > F      =  0.0475
Residual |  11448.2732    94  121.790141            R-squared     =  0.1107
---------+------------------------------           Adj R-squared =  0.0634
   Total |   12873.36    99  130.033939            Root MSE      =  11.036

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
     sex |   .8919085   2.248562     0.397   0.693    -3.572663    5.35648
     tox |  -1.956994   2.991707    -0.654   0.515    -7.897098    3.98311
  grmhem |   -4.17651   3.252088    -1.284   0.202    -10.63361   2.280587
 gestage |   1.262544   .4855007     2.600   0.011       .298571   2.226517
  apgar5 |   .3452043   .4766998     0.724   0.471    -.6012943   1.291703
   _cons |   9.092581    13.888     0.655   0.514    -18.48237   36.66753
------------------------------------------------------------------------------
```

それでは妊娠週数と中毒症を掛け合わせたパラメーターを創ってみましょう。

```
. generate gestox=gestage*tox
```

これを交えて解析します。

```
. regress  sbp sex tox grmhem gestage apgar5 gestox

  Source |       SS       df       MS              Number of obs =     100
---------+------------------------------           F(  6,    93) =    2.10
   Model |  1534.70241     6  255.783735            Prob > F      =  0.0608
Residual |  11338.6576    93  121.921049            R-squared     =  0.1192
---------+------------------------------           Adj R-squared =  0.0624
   Total |   12873.36    99  130.033939            Root MSE      =  11.042
```

```
------------------------------------------------------------------------------
       sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
       sex |    .672851   2.261601     0.298   0.767    -3.818239    5.163941
       tox |   32.74528   36.72051     0.892   0.375    -40.17438    105.6649
     grmhem |  -4.037246   3.257149    -1.240   0.218     -10.5053    2.430805
   gestage |   1.490856   .5421644     2.750   0.007     .4142248    2.567487
     apgar5 |   .3863009   .4789212     0.807   0.422    -.5647416    1.337343
     gestox |  -1.141672   1.204049    -0.948   0.345    -3.532675    1.249332
      _cons |   2.437269   15.56758     0.157   0.876    -28.47685    33.35139
------------------------------------------------------------------------------
```

今まで諸々の因子を加えて相関関係がどのように変わるか見てきましたが、結局妊娠週数のみ p=0.007 で有意であり、他は雑音(noise)です。

Model Selection

もしも上のようにcoefficient β_iが 0 と違わない（有意差がなくH_0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ます。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すが、これは統計的および非統計的判断により決定します。サンプル数に比べてあまりにも変数が多いと値の動揺が激しくなります。まず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しなくてはなりません。例えば低出生体重児の血圧を考えた時、何が影響しそうかのアイデ$_{ィ}$アが無いと話がはじまりません。

まず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します。これをall possible models と呼びます。変数が少ないときは何とかなるかもしれませんが、変数が多いときは不可能です。そこで我々は 2 種類のstepwise approach を用います。１つはforward selection で、我々は最も大きなcoefficient をもつ変数から始め、次に R^2 を最も大きくした変数を加えます。そして変化を認めなくなった時点で終わります。一方backward eliminationでは、最初に有意な変数を全て加えておいて１つずつ減らしていきます。Forward selection の逆ですから、R^2がほとんど変化がない場合その変数を抜くことができます。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ます。

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すが、我々が注意しなくてはならないのは collinearlity です。Collinearity は２つ以上の変数がお互い連動するとき発生します。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ますが、車の交通量の多寡は大気汚染の程度に直接影響してきますから、collinearlity

を生じます。Collinearlity を生じると coefficient and/or standard error の変動が大きくなります。例えば

	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

この表ではx_2を加えることによってR^2は変化していないのにcoefficient は倍に、standard error は 10 倍になっています。このように特にstandard error が大きく変動する場合にはcollinearity を考慮しなくてはなりません。このような場合x_1とx_2の間のcorrelation を調べてみればすぐにわかることです。このような場合しばしばx_iはx_2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います。よってcollinearityにあたる変数を加えてはいけません。