STATAnote_log transformation

Log transformation

ความเบ้ (skew) ของ Predictor เป็นภัยต่อ linearity assumption (ซึ่งต้องมีใน linear และ logistic regression)
ความเบ้ (skew) ของ Outcome เป็นอนั ตรายตอ่ equal varience assumption (ซึ่งตอ้ งมีใน linear regression)

ยกตัวอย่างข้อมูลสมมติ age เป็น predictor ของ Triglyceride level

Concept 1: log คือ natural log
.การ transform เมื่อใช้ function "log" เฉยๆ จะหมายถึง natural log (ln) ซึ่งในการวิเคราะห์ด้วย STATA เป็นที่นิยมกว่า

```
. gen lnage = log(age)
(4 missing values generated)
```

ถา้ จะ transform เป็น log ฐาน 10 ตอ้ ง

```
.gen lo10age = log10(age)
. gen lntgl = log(tgl)
(4 missing values generated)
. gen log10tgl = log10(tgl)
(4 missing values generated)
. sum tgl
Variable | Obs   Mean    Std. Dev.   Min      Max
-------------+--------------------------------------------------------
     tgl | 2759 166.1493 63.51077    31      476

.sum lntgl
Variable | Obs  Mean      Std.Dev.  Min       Max
-------------+--------------------------------------------------------
lntgl    | 2759 5.036282 .4010721 3.433987 6.165418

.sum log10tg
Variable | Obs    Mean     Std. Dev. Min      Max
-------------+--------------------------------------------------------
log10tgl | 2759 2.187229 .1741834 1.491362 2.677607
```

****

Concep 2: log transform ของ outcome หาผลต่าง coef ด้วย exp form
ทบทวนหลักคณิตศาสตร์

log(B1) - log(B0) = log(B1 /B0) -> นีคือ beta หรือ coef different
ดังนั้น B1/B0 = exp (beta)
Percentage difference = 100 * (B1/B0 -1.00) = ส่วนที่ต่างจาก 1.00 ของ exp(beta)เป็น %

กรณีใช้วิธีปกติ

```
. regress logtgl age

      Source |       SS       df       MS              Number of obs =     216
-------------+------------------------------           F(  1,   214) =   62.62
       Model |  .201379702     1  .201379702           Prob > F      =  0.0000
    Residual |  .688250499   214  .003216124           R-squared     =  0.2264
-------------+------------------------------           Adj R-squared =  0.2227
       Total |    .8896302   215  .004137815           Root MSE      =  .05671

------------------------------------------------------------------------------
      logtgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0043552   .0005504     7.91   0.000     .0032703      .00544
       _cons |   3.772187   .0373861   100.90   0.000     3.698495    3.845879
------------------------------------------------------------------------------
```

การแปลผล ต้องไม่ลืมวา coef คือผลต่างของ log ค่าจริงๆ จึงต้อง antilog ด้วย exponetial อีกที

```
. disp exp(.0043552)
1.0043647
```

เพื่อลดความยุ่งยากและกันลืม เมื่อมี log ของ outcome ควรใส่ option eform("exp(beta)") แสดง exp มาเลย

```
. regress logtgl age ,eform("exp(beta)")

      Source |       SS       df       MS              Number of obs =     216
-------------+------------------------------           F(  1,   214) =   62.62
       Model |  .201379702     1  .201379702           Prob > F      =  0.0000
    Residual |  .688250499   214  .003216124           R-squared     =  0.2264
-------------+------------------------------           Adj R-squared =  0.2227
       Total |    .8896302   215  .004137815           Root MSE      =  .05671

------------------------------------------------------------------------------
      logtgl |  exp(beta)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.004365   .0005528     7.91   0.000     1.003276    1.005455
------------------------------------------------------------------------------
```

อ่านว่า triglyceride มี relative fold 1.004 หรือ percentage increase 0.4
หรืออ่านง่ายๆ ว่า เพิ่มขึ้น 0.4% ในแต่ละ 1 ปีของอายุ
หรืออีกนัยหนึ่ง คือ เพิ่มขึ้น 4% ในแต่ละ "10 ปี" ของอายุ

มีเทคนิคการแสดงผลได้หลายแบบ ที่ให้ผลแบบเดียวกับคิดข้างต้น

```
. lincom age*10, eform

 ( 1)  10*age = 0

------------------------------------------------------------------------------
      logtgl |     exp(b)   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.044514   .0057488     7.91   0.000     1.033244    1.055907
------------------------------------------------------------------------------

. nlcom 100*(exp(_b[age]*10)-1)

       _nl_1:  100*(exp(_b[age]*10)-1)

------------------------------------------------------------------------------
      logtgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   4.451399   .5748808     7.74   0.000     3.318245    5.584553
------------------------------------------------------------------------------
```

****

Concept 3: log transformation ของ predictor : "for each 1% increse unit predictor"

สูตร coef * ln(1.01) = mean outcome per 1 * increase unit predictor

ตัวอย่าง

```
. regress tgl lnage bmi

      Source |       SS       df       MS              Number of obs =     216
-------------+------------------------------           F(  2,   213) =   66.36
       Model |  1173.14615     2  586.573075           Prob > F      =  0.0000
    Residual |  1882.81425   213   8.8395035           R-squared     =  0.3839
-------------+------------------------------           Adj R-squared =  0.3781
```

```
       Total |   3055.9604   215  14.2137693           Root MSE      =  2.9731

------------------------------------------------------------------------------
         tgl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnage |   15.10638   1.862788     8.11   0.000     11.43452    18.77824
         bmi |  -.2878141    .038252    -7.52   0.000     -.363215   -.2124131
       _cons |   2.728132   7.988542     0.34   0.733    -13.01859    18.47486
------------------------------------------------------------------------------
```

อ่านว่า mean triglyceride increae 0.15 mg/dl ต่อ 1% year increase age
หรืออีกนัยหนึ่งคือ 1.5 mg/dl ต่อ 10% year increase age.

```
. disp (15.1* ln(1.01))
.15025


 closed on:  10 Mar 2011, 23:00:03
-------------------------------------------------------------------------------
-------------------------------------------------
```