Intermediate Stata

Hsueh-Sheng Wu

CFDR Workshop Series

Spring 2011

BGSU
Center for Family and Demographic Research

Outline

- Validity and Reliability of A Scale
  - Different types of reliability and validity
  - Cronbach's Alpha (Inter-item Reliability)
  - Factor Analysis (Construct Validity)
- Inference Analyses for Continuous Dependent Variables
  - T-test
  - Anova
  - Ordinary Least Squares (OLS) regression
- Inference Analysis for Categorical Dependent Variables
  - Chi-square test
  - Testing proportions
  - Logistic regression
  - Ordered logistic regression
  - Multinomial logistic regression

Validity and Reliability of A Scale

- Reliability:
  - Inter-item reliability (internal consistency)
  - Inter-observer reliability
  - Test-retest reliability
  - Alternate-forms reliability (split-halves reliability)

- Validity:
  - Face validity
  - Criterion validity: concurrent validity and predictive validity
  - Construct validity: convergent validity and discriminant validity
  - Content validity

Cronbach's Alpha (Inter-Item Reliability)

- Stata example:

. webuse bg2
. describe

```
Contains data from http://www.stata-press.com/data/r11/bg2.dta
  obs:           568                          Physician-cost data
 vars:             7                          11 Feb 2009 21:54
 size:        17,040 (99.9% of memory free)   (_dta has notes)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label    variable label
-------------------------------------------------------------------------------
clinid          int    %9.0g                Physician identifier
bg2cost1        float  %9.0g                Best health care is expensive
bg2cost2        float  %9.0g                Cost is a major consideration
bg2cost3        float  %9.0g                Determine cost of tests first
bg2cost4        float  %9.0g                Monitor likely complications only
bg2cost5        float  %9.0g                Use all means regardless of cost
bg2cost6        float  %9.0g                Prefer unnecessary tests to missing tests
-------------------------------------------------------------------------------
Sorted by:  clinid
```

Cronbach's Alpha (Continued)

- **pwcorr bg2cost1-bg2cost6, star(.05)**

```
             | bg2cost1 bg2cost2 bg2cost3 bg2cost4 bg2cost5 bg2cost6
-------------+------------------------------------------------------
    bg2cost1 |   1.0000
    bg2cost2 |   0.0920*  1.0000
    bg2cost3 |   0.0540   0.3282*  1.0000
    bg2cost4 |  -0.0380   0.1420*  0.2676*  1.0000
    bg2cost5 |   0.2380* -0.1394* -0.0550  -0.0567   1.0000
    bg2cost6 |   0.2431* -0.0671  -0.1075* -0.1329*  0.3524*  1.0000
```

- **alpha bg2cost1-bg2cost6**

```
Test scale = mean(unstandardized items)
Reversed items:  bg2cost2 bg2cost3 bg2cost4

Average interitem covariance:      .134797
Number of items in the scale:            6
Scale reliability coefficient:      0.4831
```

Center for Family and Demographic Research

Cronbach's Alpha (Continued)

```
. alpha bg2cost1 bg2cost2 bg2cost3 bg2cost4 bg2cost5 bg2cost6, item

Test scale = mean(unstandardized items)

                                                          average
                          item-test     item-rest       interitem
Item          |  Obs  Sign   correlation   correlation     covariance      alpha
--------------+-----------------------------------------------------------------
bg2cost1      |  568    +       0.4333        0.1296        .1648775      0.4968
bg2cost2      |  568    -       0.5000        0.2083        .1437329      0.4563
bg2cost3      |  568    -       0.5378        0.2549        .1317608      0.4314
bg2cost4      |  568    -       0.5166        0.2286        .1384659      0.4456
bg2cost5      |  568    +       0.5810        0.3101        .1180535      0.4009
bg2cost6      |  568    +       0.6005        0.3357        .1118917      0.3865
--------------+-----------------------------------------------------------------
Test scale    |                                             .134797      0.4831
```

BGSU Center for Family and Demographic Research

Cronbach's Alpha (Continued)

```
. alpha bg2cost2 bg2cost3 bg2cost4 bg2cost5 bg2cost6
Test scale = mean(unstandardized items)
Reversed items:  bg2cost5 bg2cost6

Average interitem covariance:     .1648775
Number of items in the scale:            5
Scale reliability coefficient:      0.4968
```

```
. alpha bg2cost2 bg2cost3 bg2cost4 bg2cost5 bg2cost6,
    item
Test scale = mean(unstandardized items)
-                                                          average
                          item-test     item-rest       interitem
Item         |  Obs  Sign   correlation   correlation     covariance      alpha
-------------+------------------------------------------------------------------
bg2cost2     |  568    +       0.5821        0.2775        .1620242      0.4361
bg2cost3     |  568    +       0.6104        0.3154        .1484126      0.4108
bg2cost4     |  568    +       0.5552        0.2427        .1749139      0.4589
bg2cost5     |  568    -       0.5566        0.2445        .1742236      0.4577
bg2cost6     |  568    -       0.5762        0.2699        .1648135      0.4411
-------------+------------------------------------------------------------------
Test scale   |                                             .1648775      0.4968
```

Center for Family and Demographic Research

Factor Analysis (Construct Validity)

. factor bg2cost1-bg2cost6, ml

```
Factor analysis/correlation                       Number of obs    =        568
Method: principal factors                         Retained factors =          3
Rotation: (unrotated)                             Number of params =         15

    --------------------------------------------------------------------------
        Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    ------------+-------------------------------------------------------------
        Factor1 |      0.85389      0.31282            1.0310       1.0310
        Factor2 |      0.54107      0.51786            0.6533       1.6844
        Factor3 |      0.02321      0.17288            0.0280       1.7124
        Factor4 |     -0.14967      0.03951           -0.1807       1.5317
        Factor5 |     -0.18918      0.06197           -0.2284       1.3033
        Factor6 |     -0.25115            .           -0.3033       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  269.07 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    ------------------------------------------------------------
        Variable |  Factor1   Factor2   Factor3 |   Uniqueness
    -------------+------------------------------+--------------
        bg2cost1 |   0.2470    0.3670   -0.0446 |      0.8023
        bg2cost2 |  -0.3374    0.3321   -0.0772 |      0.7699
        bg2cost3 |  -0.3764    0.3756    0.0204 |      0.7169
        bg2cost4 |  -0.3221    0.1942    0.1034 |      0.8479
        bg2cost5 |   0.4550    0.2479    0.0641 |      0.7274
        bg2cost6 |   0.4760    0.2364   -0.0068 |      0.7175
    ------------------------------------------------------------
```

Factor Analysis (Continued)

```
. factor bg2cost1-bg2cost6, factor(2) ml

Factor analysis/correlation                       Number of obs    =        568
    Method: maximum likelihood                    Retained factors =          2
    Rotation: (unrotated)                         Number of params =         11

--------------------------------------------------------------------------
    Factor  |   Eigenvalue   Difference        Proportion   Cumulative
------------+-------------------------------------------------------------
    Factor1 |      1.02766      0.28115            0.5792       0.5792
    Factor2 |      0.74651            .            0.4208       1.0000
--------------------------------------------------------------------------
LR test: independent vs. saturated:  chi2(15) =  269.07 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    ---------------------------------------------
    Variable |  Factor1   Factor2 |   Uniqueness
    ---------+--------------------+--------------
    bg2cost1 |  -0.1371    0.4235 |      0.8018
    bg2cost2 |   0.4140    0.1994 |      0.7888
    bg2cost3 |   0.6199    0.3692 |      0.4794
    bg2cost4 |   0.3577    0.0909 |      0.8638
    bg2cost5 |  -0.3752    0.4355 |      0.6695
    bg2cost6 |  -0.4295    0.4395 |      0.6224
```

Factor Analysis (Continued)

```
. rotate, varimax
Factor analysis/correlation                       Number of obs    =      568
    Method: maximum likelihood                    Retained factors =        2
    Rotation: orthogonal varimax (Kaiser off)     Number of params =       11
--------------------------------------------------------------------------
        Factor  |     Variance   Difference        Proportion   Cumulative
    ------------+-------------------------------------------------------------
       Factor1  |      0.89639      0.01863            0.5052       0.5052
       Factor2  |      0.87777            .            0.4948       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(15) =  269.07 Prob>chi2 = 0.0000
Rotated factor loadings (pattern matrix) and unique variances
    -----------------------------------------------
       Variable |  Factor1   Factor2 |   Uniqueness
    ------------+--------------------+--------------
       bg2cost1 |   0.4276    0.1239 |      0.8018
       bg2cost2 |  -0.0670    0.4547 |      0.7888
       bg2cost3 |  -0.0417    0.7203 |      0.4794
       bg2cost4 |  -0.1253    0.3471 |      0.8638
       bg2cost5 |   0.5710   -0.0666 |      0.6695
       bg2cost6 |   0.6047   -0.1093 |      0.6224
    -----------------------------------------------
Factor rotation matrix
    --------------------------------
                 | Factor1  Factor2
    -------------+------------------
       Factor1 | -0.5606
       Factor2 |  0.8281   0.5606
    --------------------------------
```

Inference Analysis (Continuous Dependent Variables)

- t-test
- ANOVA
- Ordinary Least Squares (OLS) regression

t-test

. webuse auto

. sum mpg foreign

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         mpg |        74     21.2973    5.785503         12         41
     foreign |        74    .2972973    .4601885          0          1
```

. ttest mpg, by(foreign)

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      52    19.82692     .657777    4.743297    18.50638    21.14747
 Foreign |      22    24.77273     1.40951    6.611187    21.84149    27.70396
---------+--------------------------------------------------------------------
combined |      74     21.2973    .6725511    5.785503     19.9569    22.63769
---------+--------------------------------------------------------------------
    diff |           -4.945804    1.362162               -7.661225   -2.230384
------------------------------------------------------------------------------
    diff = mean(Domestic) - mean(Foreign)                         t =  -3.6308
Ho: diff = 0                                     degrees of freedom =       72

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0003         Pr(|T| > |t|) = 0.0005          Pr(T > t) = 0.9997
```

ANOVA

- **anova mpg foreign**

```
                   Number of obs =      74     R-squared     =  0.1548
                   Root MSE      = 5.35582     Adj R-squared =  0.1430

    Source |  Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------
     Model |  378.153515     1  378.153515      13.18     0.0005
           |
   foreign |  378.153515     1  378.153515      13.18     0.0005
           |
  Residual |  2065.30594    72  28.6848048
-----------+----------------------------------------------------
     Total |  2443.45946    73  33.4720474
```

OLS Regression

- **reg mpg foreign**

```
      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,    72) =   13.18
       Model |  378.153515     1  378.153515           Prob > F      =  0.0005
    Residual |  2065.30594    72  28.6848048           R-squared     =  0.1548
-------------+------------------------------           Adj R-squared =  0.1430
       Total |  2443.45946    73  33.4720474           Root MSE      =  5.3558

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   4.945804   1.362162     3.63   0.001     2.230384    7.661225
       _cons |   19.82692   .7427186    26.70   0.000     18.34634    21.30751
------------------------------------------------------------------------------
```

Inference Analysis (Categorical Dependent Variables)

- Chi-square test
- Testing proportions
- Logistic regression
- Ordered logistic regression
- Multinomial logistic regression

Chi-square Test

```
. webuse auto
. label define repair 1 "very poor" 2 "poor" 3 "average" 4 "good" 5 "very
    good"
. label values rep78 repair
. tab rep78 foreign, chi2
```

Repair Record 1978	Car type: Domestic	Foreign	Total
very poor	2	0	2
poor	8	0	8
average	27	3	30
good	9	9	18
very good	2	9	11
Total	48	21	69

Pearson chi2(4) = 27.2640 Pr = 0.000

Testing Proportions

```
. gen r_price=0 if price<10000
. replace r_price=1 if price >=10000
. tab r_price foreign, col
```

r_price	Car type: Domestic	Foreign	Total
< $10,000	44	20	64
	84.62	90.91	86.49
>= $10,000	8	2	10
	15.38	9.09	13.51
Total	52	22	74
	100.00	100.00	100.00

BGSU Center for Family and Demographic Research

Testing Proportions (Continued)

. prtest r_price, by(foreign)

```
Two-sample test of proportion                Domestic: Number of obs =       52
                                              Foreign: Number of obs =       22
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Domestic |   .1538462   .0500341                      .0557811    .2519112
     Foreign |   .0909091   .0612909                     -.0292189     .211037
-------------+----------------------------------------------------------------
        diff |   .0629371   .0791201                     -.0921355    .2180096
             |  under Ho:   .0869483     0.72   0.469
------------------------------------------------------------------------------
        diff = prop(Domestic) - prop(Foreign)                     z =   0.7238
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.7654         Pr(|Z| < |z|) = 0.4692          Pr(Z > z) = 0.2346
```

Logistic Regression

- **logit r_price foreign**

```
Iteration 0:   log likelihood = -29.306449
Iteration 1:   log likelihood = -29.029957
Iteration 2:   log likelihood = -29.026792
Iteration 3:   log likelihood = -29.026791

Logistic regression                                     Number of obs   =         74
                                                        LR chi2(1)      =       0.56
                                                        Prob > chi2     =     0.4545
Log likelihood = -29.026791                             Pseudo R2       =     0.0095

------------------------------------------------------------------------------
     r_price |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   -.597837   .8353007    -0.72   0.474    -2.234996    1.039322
       _cons |  -1.704748   .3843531    -4.44   0.000    -2.458066   -.9514299
------------------------------------------------------------------------------
```

Ordered Logistic Regression

. ologit rep78 foreign

```
Iteration 0:   log likelihood = -93.692061
Iteration 1:   log likelihood = -79.696089
Iteration 2:   log likelihood = -79.034005
Iteration 3:   log likelihood = -79.029244
Iteration 4:   log likelihood = -79.029243

Ordered logistic regression                       Number of obs   =         69
                                                  LR chi2(1)      =      29.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -79.029243                       Pseudo R2       =     0.1565

------------------------------------------------------------------------------
       rep78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |    2.98155   .6203644     4.81   0.000     1.765658    4.197442
-------------+----------------------------------------------------------------
       /cut1 |  -3.158382   .7224269                     -4.574313   -1.742452
       /cut2 |  -1.362642   .3557343                     -2.059868   -.6654154
       /cut3 |   1.232161   .3431227                      .5596532    1.90467
       /cut4 |   3.246209   .5556657                      2.157124    4.335293
------------------------------------------------------------------------------
```

BGSU Center for Family and Demographic Research

Multinomial Logistic Regression

. webuse sysdsn1
. describe

```
Contains data from http://www.stata-press.com/data/r11/sysdsn1.dta
  obs:           644                          Health insurance data
 vars:            13                          28 Mar 2009 13:10
 size:        16,744 (99.9% of memory free)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
patid           float  %9.0g
age             float  %10.0g                 NEMC (ISCNRD-IBIRTHD)/365.25
male            byte   %8.0g                  NEMC PATIENT MALE
nonwhite        float  %9.0g
insure          byte   %9.0g       insure
------------------------------------------------------------------------------
Sorted by:  patid
```

. tab insure

```
     insure |      Freq.     Percent        Cum.
------------+-----------------------------------
  Indemnity |        294       47.73       47.73
    Prepaid |        277       44.97       92.69
   Uninsure |         45        7.31      100.00
------------+-----------------------------------
      Total |        616      100.00
```

Multinomial Logistic Regression (Cont.)

. mlogit insure age male nonwhite

```
Iteration 0:   log likelihood = -555.85446
Iteration 1:   log likelihood = -545.60089
Iteration 2:   log likelihood = -545.58328
Iteration 3:   log likelihood = -545.58328
Multinomial logistic regression                 Number of obs   =        615
                                                LR chi2(6)      =      20.54
                                                Prob > chi2     =     0.0022
Log likelihood = -545.58328                     Pseudo R2       =     0.0185

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Indemnity    |  (base outcome)
-------------+----------------------------------------------------------------
Prepaid      |
         age |  -.0111915   .0060915    -1.84   0.066    -.0231305    .0007475
        male |   .5739825   .2005221     2.86   0.004     .1809665    .9669985
    nonwhite |   .7312659    .218978     3.34   0.001      .302077    1.160455
       _cons |   .1567003   .2828509     0.55   0.580    -.3976773    .7110778
-------------+----------------------------------------------------------------
Uninsure     |
         age |  -.0058414   .0114114    -0.51   0.609    -.0282073    .0165245
        male |   .5102237   .3639793     1.40   0.161    -.2031626     1.22361
    nonwhite |   .4333141   .4106255     1.06   0.291     -.371497    1.238125
       _cons |  -1.811165   .5348606    -3.39   0.001    -2.859473   -.7628578
------------------------------------------------------------------------------
```

Conclusion

- Among different types of reliability and validity, only Inter-item Reliability and Construct validity can be directly tested without using additional data.
- Confirmatory factor analysis is need for truly testing construct validity, which you need to use Structural Equation Software (e.g., SAS, LISREL, M-Plus) to do.
- The measurement of the dependent variables contributes to what analyses to be conducted.
- Try to understand what different options in the Stata commands.
- Other useful resources for learning using Stata.
  - http://www.ats.ucla.edu/stat/stata/
  - http://data.princeton.edu/stata/
  - http://www.cpc.unc.edu/research/tools/data_analysis/statatutorial/index.html