. do "C:\Users\carmelia\Desktop\stndzxage\ado\stndzxage tutorial.do"

. *stndzxage tutorial
. *by Sarah Reynolds
. *2-27-19

. *The file checks how the command stndzxage differs from zscore
. *The file illustrates how to use the command

. clear all

. set more off

. cd "C:\Users\carmelia\Desktop\stndzxage\ado"
C:\Users\carmelia\Desktop\stndzxage\ado

. use "stndzxage_sample_data.dta", clear

. count
  1,429

. *1,429 children in the data

. count if TestScore~=.
  1,420

. *1,420 were tested

. hist AgeMonth
 (bin=31, start=10, width=.58064516)
. *ages concentrated in the center

. stndzxage TestScore AgeMonth

. sum stx_TestScore

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>stx_TestScore</td>
<td>1,332</td>
<td>4.73e-08</td>
<td>.9954819</td>
<td>-3.847551</td>
<td>3.815243</td>
</tr>
</tbody>
</table>

. *mean about 0 & standard deviation about 1, as expected
. *however, there are fewer observations

. *Do a loop to check standardization with stata command
. levelsof AgeMonth, local(ages)
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 28

. gen Z_TestScore=.
(1,429 missing values generated)

. foreach age of local ages {
2.    zscore TestScore if AgeMonth==`age'
3.    replace Z_TestScore=z_TestScore if AgeMonth==`age'
4.    drop z_TestScore
5. }

z_TestScore created with 1410 missing values
(19 real changes made)
z TestScore created with 1401 missing values
(28 real changes made)
z_TestScore created with 1373 missing values
(56 real changes made)
z TestScore created with 1350 missing values
(79 real changes made)
z_TestScore created with 1316 missing values  
(113 real changes made)  
z_TestScore created with 1320 missing values  
(109 real changes made)  
z_TestScore created with 1304 missing values  
(125 real changes made)  
z_TestScore created with 1324 missing values  
(105 real changes made)  
z_TestScore created with 1301 missing values  
(128 real changes made)  
z_TestScore created with 1303 missing values  
(126 real changes made)  
z_TestScore created with 1309 missing values  
(120 real changes made)  
z_TestScore created with 1317 missing values  
(112 real changes made)  
z_TestScore created with 1318 missing values  
(111 real changes made)  
z_TestScore created with 1351 missing values  
(78 real changes made)  
z_TestScore created with 1359 missing values  
(70 real changes made)  
z_TestScore created with 1401 missing values  
(28 real changes made)  
z_TestScore created with 1417 missing values  
(12 real changes made)  
z_TestScore created with 1429 missing values  
(0 real changes made)  

. sum Z_TestScore

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z_TestScore</td>
<td>1,419</td>
<td>9.17e-10</td>
<td>.9943422</td>
<td>-3.847551</td>
<td>3.815243</td>
</tr>
</tbody>
</table>

*mean about 0 & standard deviation about 1, as expected  
*however, there are more observations, equal to the  
*number of children who took the test - 1

. sum AgeMonth if Z_TestScore==. & TestScore==.  

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeMonth</td>
<td>1</td>
<td>28</td>
<td>.</td>
<td>28</td>
<td>28</td>
</tr>
</tbody>
</table>

*The - 1 corresponds to the child who was the only one of thier age  
*Check to see how well they line up if there are both standardization variables  
.scatter Z_TestScore stx_TestScore
. tab AgeMonth if stx_TestScore~=.

<table>
<thead>
<tr>
<th>AgeMonth</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>56</td>
<td>4.20</td>
<td>4.20</td>
</tr>
<tr>
<td>13</td>
<td>79</td>
<td>5.93</td>
<td>10.14</td>
</tr>
<tr>
<td>14</td>
<td>113</td>
<td>8.48</td>
<td>18.62</td>
</tr>
<tr>
<td>15</td>
<td>109</td>
<td>8.18</td>
<td>26.80</td>
</tr>
<tr>
<td>16</td>
<td>125</td>
<td>9.38</td>
<td>36.19</td>
</tr>
<tr>
<td>17</td>
<td>105</td>
<td>7.88</td>
<td>44.07</td>
</tr>
<tr>
<td>18</td>
<td>128</td>
<td>9.61</td>
<td>53.68</td>
</tr>
<tr>
<td>19</td>
<td>126</td>
<td>9.46</td>
<td>63.14</td>
</tr>
<tr>
<td>20</td>
<td>120</td>
<td>9.01</td>
<td>72.15</td>
</tr>
<tr>
<td>21</td>
<td>112</td>
<td>8.41</td>
<td>80.56</td>
</tr>
<tr>
<td>22</td>
<td>111</td>
<td>8.33</td>
<td>88.89</td>
</tr>
<tr>
<td>23</td>
<td>78</td>
<td>5.86</td>
<td>94.74</td>
</tr>
<tr>
<td>24</td>
<td>70</td>
<td>5.26</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 1,332 | 100.00  

. tab AgeMonth if Z_TestScore~=.

<table>
<thead>
<tr>
<th>AgeMonth</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>19</td>
<td>1.34</td>
<td>1.34</td>
</tr>
<tr>
<td>11</td>
<td>28</td>
<td>1.97</td>
<td>3.31</td>
</tr>
<tr>
<td>12</td>
<td>56</td>
<td>3.95</td>
<td>7.26</td>
</tr>
<tr>
<td>13</td>
<td>79</td>
<td>5.57</td>
<td>12.83</td>
</tr>
</tbody>
</table>
14 | 113 | 7.96 | 20.79
15 | 109 | 7.68 | 28.47
16 | 125 | 8.81 | 37.28
17 | 105 | 7.40 | 44.68
18 | 128 | 9.02 | 53.70
19 | 126 | 8.88 | 62.58
20 | 120 | 8.46 | 71.04
21 | 112 | 7.89 | 78.93
22 | 111 | 7.82 | 86.75
23 | 78  | 5.50 | 92.25
24 | 70  | 4.93 | 97.18
25 | 28  | 1.97 | 99.15
26 | 12  | 0.85 | 100.00

Total | 1,419 | 100.00

*mismatch in missings because stndzxage has 30 observations minimum

*find out how many are in each month to re-standardize the
*using the smallest number of observations!

.tab AgeMonth

AgeMonth | Freq. | Percent | Cum.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>19</td>
<td>1.33</td>
<td>1.33</td>
</tr>
<tr>
<td>11</td>
<td>28</td>
<td>1.96</td>
<td>3.29</td>
</tr>
<tr>
<td>12</td>
<td>58</td>
<td>4.06</td>
<td>7.35</td>
</tr>
<tr>
<td>13</td>
<td>79</td>
<td>5.53</td>
<td>12.88</td>
</tr>
<tr>
<td>14</td>
<td>114</td>
<td>7.98</td>
<td>20.85</td>
</tr>
<tr>
<td>15</td>
<td>109</td>
<td>7.63</td>
<td>28.48</td>
</tr>
<tr>
<td>16</td>
<td>125</td>
<td>8.75</td>
<td>37.23</td>
</tr>
<tr>
<td>17</td>
<td>106</td>
<td>7.42</td>
<td>44.65</td>
</tr>
<tr>
<td>18</td>
<td>130</td>
<td>9.10</td>
<td>53.74</td>
</tr>
<tr>
<td>19</td>
<td>127</td>
<td>8.89</td>
<td>62.63</td>
</tr>
<tr>
<td>20</td>
<td>120</td>
<td>8.40</td>
<td>71.03</td>
</tr>
<tr>
<td>21</td>
<td>113</td>
<td>7.91</td>
<td>78.94</td>
</tr>
<tr>
<td>22</td>
<td>112</td>
<td>7.84</td>
<td>86.77</td>
</tr>
<tr>
<td>23</td>
<td>78</td>
<td>5.46</td>
<td>92.23</td>
</tr>
<tr>
<td>24</td>
<td>70</td>
<td>4.90</td>
<td>97.13</td>
</tr>
<tr>
<td>25</td>
<td>28</td>
<td>1.96</td>
<td>99.09</td>
</tr>
<tr>
<td>26</td>
<td>12</td>
<td>0.84</td>
<td>99.93</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>0.07</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 1,429 | 100.00

.stndzxage TestScore AgeMonth, minbinsize(12)

. assert stx_TestScore==Z_TestScore
1,299 contradictions in 1,429 observations
assertion is false
r(9);

eod of do-file


```
r(9);
.do "C:\Users\carmelia\AppData\Local\Temp\STD00000000.tmp"

. *This error turns out to be from rounding
. gen stx_round=round(stx_TestScore, 0.0001)
(10 missing values generated)

. gen Z_round=round(Z_TestScore, 0.0001)
(10 missing values generated)

. assert stx_round==Z_round

. *****Validation complete********
.
.
. ****Exploring options****
.
. *GRAPHING
. stndzxage TestScore AgeMonth, graph
(9 observations deleted)

![Raw & Standardized Scores by Age (integer)](image)

Raw shows mean (or median) used to standardize.
Standardized shows mean of standardized scores.

. *Notice there are more ages with raw data points than have means
. *These ages had too few observations (default minbinsize is 30)

. *BIN WIDTH
. *let's widen the age bins so more ages are grouped together, resulting in
. *a larger number of observations in each bin
. stndzxage TestScore AgeMonth, binwidth(6) graph
(9 observations deleted)
```
*the waves in the standardized data indicate bins are probably too wide
.stndxage TestScore AgeMonth, binwidth(3) graph
(9 observations deleted)

*still some age dependency but not so much
*note the last bin included 4 ages (see help file chart about bin grouping)

*MINIMUM BIN SIZE
*let's increase the minimum number of observations allowed in each bin
.stndxage TestScore AgeMonth, binwidth(3) minbinsize(150) graph
(9 observations deleted)
. *CONTINUOUS
. *continuous standardization is a good option when data density has gaps (in tails)
. stndzTestScore AgeMonth, continuous graph
   (9 observations deleted)

. sum stx_TestScore

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>stx_TestScore</td>
<td>1,420</td>
<td>.0001268</td>
<td>1.001219</td>
<td>-4.085064</td>
<td>3.63665</td>
</tr>
</tbody>
</table>

. *note all observations are standardized
. stndzTestScore AgeMonth, continuous poly(1) graph // linear
   (9 observations deleted)
*STANDARDIZING OVER ADDITIONAL VARIABLES

*you can use if to standardize a single subgroup

```
stndzxage TestScore AgeMonth, continuous poly(5) graph // a bit more curvature
(9 observations deleted)
```

```
<p>| Summary of Stndz TestScore by AgeMonth ~N(0,1) binwidth 3 |
|-------------|-------------|-------------|-------------|</p>
<table>
<thead>
<tr>
<th>Male</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>-7.448e-09</td>
<td>.99707603</td>
<td>686</td>
</tr>
<tr>
<td>Total</td>
<td>-7.448e-09</td>
<td>.99707603</td>
<td>686</td>
</tr>
</tbody>
</table>
```

```
stndzxage TestScore AgeMonth if Male==0, binwidth(3)
```

```
<p>| Summary of Stndz TestScore by AgeMonth ~N(0,1) binwidth 3 |
|-------------|-------------|-------------|-------------|</p>
<table>
<thead>
<tr>
<th>Male</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>-7.448e-09</td>
<td>.99707603</td>
<td>686</td>
</tr>
<tr>
<td>Total</td>
<td>-7.448e-09</td>
<td>.99707603</td>
<td>686</td>
</tr>
</tbody>
</table>
```

```
``
Summary of Stndz TestScore by AgeMonth ~N(0,1) binwidth 3

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Freq.</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Female</td>
<td>1.224e-07</td>
<td>.99710566</td>
<td>693</td>
</tr>
<tr>
<td>Total</td>
<td>1.224e-07</td>
<td>.99710566</td>
<td>693</td>
</tr>
</tbody>
</table>

*but below is more efficient

*standardize by age & gender
stndzxage TestScore AgeMonth Male, binwidth(3) graph (9 observations deleted)

---

**Raw & Standardized Scores by Age (integer)**

Raw shows mean (or median) used to standardize. Standardized shows mean of standardized scores.

---

tab Male, sum(stx_TestScore)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Mean</td>
<td>Std. Dev.</td>
<td>Freq.</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Female</td>
<td>1.224e-07</td>
<td>.99710566</td>
<td>693</td>
</tr>
<tr>
<td>Male</td>
<td>-7.448e-09</td>
<td>.99707603</td>
<td>686</td>
</tr>
<tr>
<td>Total</td>
<td>5.780e-08</td>
<td>.99672906</td>
<td>1,379</td>
</tr>
</tbody>
</table>

*note means & s.d. are 0 in both cases

*standardize by age, gender, and urban
. stndzxage TestScore AgeMonth Male Urban, continuous graph
(9 observations deleted)

Raw & Standardized Scores by Age (integer)

Raw shows mean (or median) used to standardize. Standardized shows mean of standardized scores.

. tab Male Urban, sum(stx_TestScore)

Means, Standard Deviations and Frequencies
of Stndz TestScore by AgeMonth Male Urban ~N(0, 1)

<table>
<thead>
<tr>
<th></th>
<th>rural</th>
<th>urban</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>.00198964</td>
<td>-.00020896</td>
<td>.00066925</td>
</tr>
<tr>
<td></td>
<td>1.0023932</td>
<td>.99970047</td>
<td>1.0000767</td>
</tr>
<tr>
<td></td>
<td>286</td>
<td>430</td>
<td>716</td>
</tr>
</tbody>
</table>

Raw shows mean (or median) used to standardize. Standardized shows mean of standardized scores.
. *STANDARDIZING WITH REGARDS TO A REFERENCE GROUP
. stndzxage TestScore AgeMonth, binwidth(3) reference(Male) graph
(9 observations deleted)

. *The graph only illustrates the data for the reference group, which was used
. *for standardizing
. tab Male, sum(stx_TestScore)

|            | Summary of Stndz TestScore by AgeMonth ~N(0,1) binwidth 3; ref grp Male=1 |
|------------------------+-----------------------------------------------|
| Male                   | Mean    | Std. Dev. | Freq. |
|------------------------+-----------------------------------------------|
| Female                 | .06922886 | 1.0157441 | 693   |
| Male                   | -7.448e-09 | .99707603 | 686   |
|------------------------+-----------------------------------------------|
| Total                  | .03479013 | 1.0067312 | 1,379 |

. *note here the mean & s.d. is ~0 & ~1 for the reference group, but different for
. *the non reference group

. *USING A REFERENCE GROUP & A SUBGROUP
. *can you do it both reference group
. stndzxage TestScore AgeMonth Urban, binwidth(3) minbinsize(30) reference(Male) graph
(9 observations deleted)
Means, Standard Deviations and Frequencies of Stndz TestScore by AgeMonth Urban \~N(0,1) binwidth 3; ref grp Male=1

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Urban</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>.10767812</td>
<td>.07666479</td>
<td>.08912855</td>
</tr>
<tr>
<td></td>
<td>1.1174035</td>
<td>.95998533</td>
<td>1.0254123</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>381</td>
<td>637</td>
</tr>
<tr>
<td>Male</td>
<td>5.173e-08</td>
<td>-5.781e-08</td>
<td>-9.468e-09</td>
</tr>
<tr>
<td></td>
<td>.9946476</td>
<td>.99577762</td>
<td>.99449898</td>
</tr>
<tr>
<td></td>
<td>282</td>
<td>357</td>
<td>639</td>
</tr>
<tr>
<td>Total</td>
<td>.0512372</td>
<td>.03957895</td>
<td>.04449442</td>
</tr>
<tr>
<td></td>
<td>1.0552224</td>
<td>.97755028</td>
<td>1.0106373</td>
</tr>
<tr>
<td></td>
<td>538</td>
<td>738</td>
<td>1276</td>
</tr>
</tbody>
</table>

*USING A DIFFERENT RUNNING VARIABLE*

*Suppose the test was administered with different questions to different ages
*Cut the data at the ages for each group

. egen testgroups=cut(AgeMonth), at(10, 13, 16, 19, 25, 30)

. tostring testgroups, replace
testgroups was float now str2
. encode testgroups, gen(TestGroups)
. label values TestGroups // remove label from TestGroup2
. stndzxage TestScore TestGroups, graph
(9 observations deleted)

. rename stx_TestScore testgroups_z

. *This graph has the test groups all lumped together
. *If you want to see the ages graphed also, use the if option.
. *Select the binwidth to be the widest number of ages in a bin.
. levels TestGroups, local(groups)
1 2 3 4 5

. gen testgroups_if_z=.
(1,429 missing values generated)

. foreach i of local groups {
    2.        stndzxage TestScore AgeMonth if TestGroups==`i', binwidth(6)
    3.        replace testgroups_if_z=stx_TestScore if TestGroups==`i'
    4. }

Raw & Standardized Scores by Age (integer)

Raw shows mean (or median) used to standardize.
Standardized shows mean of standardized scores.

Raw Data
-4 -2
0 2 4
1 2 3 4 5
TestGroups
Standardized Data

Raw & Standardized Scores by Age (integer)
Raw & Standardized Scores by Age (integer)

Raw shows mean (or median) used to standardize.
Standardized shows mean of standardized scores.

(1,326 observations deleted)
(103 real changes made)

(1,128 observations deleted)
(301 real changes made)

(1,071 observations deleted)
. assert testgroups_z==testgroups_if_z

. *Though the syntax below is appealing, it does not work because
. *the ages are divided up by binwidth before the TestGroups
. * stndzxage TestScore AgeMonth TestGroups, binwidth(6) graph
. *don't use this code!
.
.
. *FLOORS & CEILINGS
. *let's make an artificial floor in this data
. replace TestScore=35 if TestScore<35
(21 real changes made)

. hist TestScore
(bin=31, start=35, width=1.2258065)
*If your data actually looked like this, you might be ok with the test ceiling, but *
you might want to rethink the appropriateness of the test for the younger kids:
*the test best discriminates after about 15 months.
*stndzxage TestScore AgeMonth, continuous graph
(9 observations deleted)
. sum stx_TestScore

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>stx_TestScore</td>
<td>1,420</td>
<td>0.0005976</td>
<td>1.002678</td>
<td>-4.073903</td>
<td>4.044211</td>
</tr>
</tbody>
</table>

. stndzxage TestScore AgeMonth, continuous floor graph
(9 observations deleted)

. sum stx_TestScore

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>stx_TestScore</td>
<td>1,420</td>
<td>0.0066415</td>
<td>0.8071882</td>
<td>-3.050389</td>
<td>2.605103</td>
</tr>
</tbody>
</table>

*The floor option uses a Tobit adjustment, which assumes a spread farther below
*that which is censored. Censoring pushes the mean up. Without the adjustment,
*the mean used to standardize is higher than the mean used to standardize with a
*Tobit adjustment. Average standardized scores are higher in the Tobit adjustment


*We can take ceilings into account as well.
replace TestScore=60 if TestScore>60 & TestScore==.
(151 real changes made)

stndzxage TestScore AgeMonth, floor ceiling minbinsize(30)
reference(Male) graph
(9 observations del
eted)

*USING THE MEDIAN & RESCALING
*The median can be used for standardizing instead of the mean.
*A different standard mean/median & standard deviation can be selected
stndzxage TestScore AgeMonth, sd(15) mean(100) binw(3)

. sum stx_TestScore

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>stx_TestScore</td>
<td>1,420</td>
<td>100</td>
<td>14.97355</td>
<td>39.86069</td>
<td>135.5473</td>
</tr>
</tbody>
</table>

. stndzxage TestScore AgeMonth, median sd(15) mean(100) binw(3)

. sum stx_TestScore

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>stx_TestScore</td>
<td>1,420</td>
<td>97.73348</td>
<td>15.05202</td>
<td>36.84235</td>
<td>133.5137</td>
</tr>
</tbody>
</table>

end of do-file

. log close
   name: <unnamed>
   log: C:\Users\carmelia\Desktop\stndzxage\ado\log2.log
   log type: text
   closed on: 7 Mar 2019, 14:05:01
---------------------------------------------------------------