Statistics/Data Analysis

```
      log:  /scratch/OhioFamilyHealthSurvey/analysislogModel.smcl
  log type:  smcl
 opened on:  16 Jul 2008, 15:54:46

1 . do /tmp/SD27137.000000

2 . /*****************************************************/
3 . /** Exploratory Data Analysis                       **/
4 . /*****************************************************/
5 .
6 . clear

7 .
8 . set memory 1g  /*increase available memory, since the data is huge*/
  (1048576k)

9 . use "ofhs_virgin_file.dta"   /*read in the data*/

10.
11. /*recode some missing data*/
12. replace s10 = . if s10>9  /*maximum recorded adults in HH = 9 */
  (0 real changes made)

13. replace s11 = . if s11>9  /*maximum recorded adults in family */
  (29 real changes made, 29 to missing)

14. replace s12 = . if s12>12 /*maximum number of children = 12 */
  (32 real changes made, 32 to missing)

15.
16.
17.
18. /*consider number of people in the family: h84 */
19. replace h84 = . if h84 == 99
  (51 real changes made, 51 to missing)

20.
21. /*consider yearly family income:  h85y*/
22. replace h85y = . if h85y > 999997
  (1440 real changes made, 1440 to missing)

23. gen sqrtincome = sqrt(h85y) /*transform to the square root scale*/
  (9938 missing values generated)

24.
25. /*consider number of hours worked: g73*/
26. replace g73 = . if g73 > 85
  (300 real changes made, 300 to missing)

27. gen hoursworked = g73
  (16754 missing values generated)

28.
29. /*consider health care premiums: b8 */
30. replace b8 = . if b8 > 99997
  (4376 real changes made, 4376 to missing)

31. gen sqrtpremiums = sqrt(b8)  /*transform to the square root scale*/
  (20846 missing values generated)
```

```
32.
33. /*b8b == 3 means the premium is paid monthly */
34.
35. /*consider the race of the adult*/
36. replace race_a = . if race_a>4
  (523 real changes made, 523 to missing)

37. /*make pretty labels*/
38. label define race 1 "White" 2 "Black" 3 "Asian" 4 "Other" 98 "Dont Know"

39. label values race_a race

40.
41. /*consider the ethnicity of the adult*/
42. replace raceth_a = . if raceth_a >4
  (0 real changes made)

43. /*make pretty labels*/
44. label define eth 1 "Hispanic" 2 "Black/NH" 3 "Asian" 4 "White/NH" 98 "Dont Know"

45. label values raceth_a eth

46.
47.
48. /*consider the age of the respondent*/
49. replace s14 = . if s14 >997
  (436 real changes made, 436 to missing)

50. gen age = s14
  (436 missing values generated)

51.
52.
53. /*generate variables for counting purposes*/
54. gen const=1

55. sum wtfinal
```

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| wtfinal | **39953** | **213.4313** | **189.372** | **3.692236** | **1944.857** |

```
56. scalar N_est = r(sum)

57. gen wtSRS = r(sum)/r(N)

58.
59.
60. /*bar plots*/
61. graph bar (sum) wtSRS, over(raceth_a) title("Equal Weights") blabel(bar) ytitle("Total")

62. graph export "barracethSRS.ps", replace
  (file barracethSRS.ps written in PostScript format)

63. graph bar (sum) wtfinal, over(raceth_a) title("Probability Weights") blabel(bar) ytitle("Total")

64. graph export "barracethWT.ps", replace
  (file barracethWT.ps written in PostScript format)

65.
```

```
66.
67. /*box plots*/
68. graph box sqrtincome if h84==1, ///
  >         ytitle("Income") title("Income for Single-Adult Households") subtitle("Equal Weights") ///
  >         ylabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20,000" ‘=sqrt(100000)’ "100,000" ‘=sqrt(300000)’ "3
  > 0000)’ "1,000,000" )

69. graph export "boxincomeSRS.ps", replace
  (file boxincomeSRS.ps written in PostScript format)

70. graph box sqrtincome if h84==1 [pweight = wtfinal], ///
  >         ytitle("Income") title("Income for Single-Adult Households") subtitle("Probability Weights
  >         ylabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20,000" ‘=sqrt(100000)’ "100,000" ‘=sqrt(300000)’ "3
  > 0000)’ "1,000,000" )

71. graph export "boxincomeWT.ps", replace
  (file boxincomeWT.ps written in PostScript format)

72.
73. /*box plots -- "zoomed" subset */
74. graph box sqrtincome if h84==1 & sqrtincome<500, ///
  >         ytitle("Income") title("Income for Single-Adult Households < 250,000") subtitle("Equal Wei
  >         ylabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20,000" ‘=sqrt(100000)’ "100,000" ‘=sqrt(300000)’ "3

75. graph export "boxincomeSRS_trunc.ps", replace
  (file boxincomeSRS_trunc.ps written in PostScript format)

76. graph box sqrtincome if h84==1  & sqrtincome<500 [pweight = wtfinal], ///
  >         ytitle("Income") title("Income for Single-Adult Households < 250,000") subtitle("Probabili
  >         ylabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20,000" ‘=sqrt(100000)’ "100,000" ‘=sqrt(300000)’ "3

77. graph export "boxincomeWT_trunc.ps", replace
  (file boxincomeWT_trunc.ps written in PostScript format)

78.
79.
80. graph box sqrtincome if h84==1 [pweight = wtfinal], ///
  >         over(cluster, label(alternate)) ///
  >         ytitle("Income") title("Income for Single-Adult Households") subtitle("Probability Weights
  >         ylabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20,000" ‘=sqrt(100000)’ "100,000" ‘=sqrt(300000)’ "3
  > 0000)’ "1,000,000" )

81. graph export "boxincomeStrata.ps", replace
  (file boxincomeStrata.ps written in PostScript format)

82.
83.
84.
85. /*scatter plots*/
86. scatter sqrtpremiums sqrtincome if h84==1, ///
  >         xlabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20,000" ‘=sqrt(100000)’ "100,000" ‘=sqrt(300000)’ "3
  > 0000)’ "1,000,000" ) ///
  >         ylabel(‘=sqrt(0)’ "0" ‘=sqrt(2500)’ "2500" ‘=sqrt(10000)’ "10,000" ‘=sqrt(22500)’ "22,500"
  >         xtitle("Income") ytitle("Premiums")

87.
88. scatter sqrtpremiums sqrtincome if h84==1 [pweight = wtfinal], ///
  >         mfcolor(none) xlabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20000" ‘=sqrt(100000)’ "100000" ‘=sqrt
  > tincome if h84==1 [pweight = wtfinal]
```

```
89.
90.
91. scatter sqrtincome hoursworked if s11==1, ///
  >         xtitle("# Hours Worked") ytitle("Income") title("Work vs. Income in Single Adult Families"
  >         ylabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20,000" ‘=sqrt(100000)’ "100,000" ‘=sqrt(300000)’ "3
  > 0000)’ "1,000,000" )

92. graph export "scatterincomehoursSRS.ps", replace
  (file scatterincomehoursSRS.ps written in PostScript format)

93. scatter sqrtincome hoursworked if h84==1 [pweight = wtfinal], ///
  >         mfcolor(none) xtitle("# Hours Worked") ytitle("Income") title("Work vs. Income in Single A
  >         ylabel(‘=sqrt(0)’ "0" ‘=sqrt(20000)’ "20,000" ‘=sqrt(100000)’ "100,000" ‘=sqrt(300000)’ "3
  > 0000)’ "1,000,000" ) ///
  >         || lfit sqrtincome hoursworked [pweight=wtfinal], lwidth(vthick)

94. graph export "scatterincomehoursWT.ps", replace
  (file scatterincomehoursWT.ps written in PostScript format)

95.
96. /*scatter plots of ordinal variables*/
97. scatter s12 s10 if region2 == "A"

98. graph export "scatterordinal1.ps", replace
  (file scatterordinal1.ps written in PostScript format)

99. scatter s12 s10 if region2 == "A" [pweight = wtfinal], mfcolor(none)

100 graph export "scatterordinal2.ps", replace
  (file scatterordinal2.ps written in PostScript format)

101 scatter s12 s10 if region2 == "A" [pweight = wtfinal], mfcolor(none) jitter(5)

102 graph export "scatterordinal3.ps", replace
  (file scatterordinal3.ps written in PostScript format)

103
104 capture bysort s12 s10 region2: egen tot_wt = total(wtfinal) /* sum the weights by the categories

105 scatter s12 s10 if region2 == "A" [pweight = tot_wt], mfcol(none)

106 graph export "scatterordinal4.ps", replace
  (file scatterordinal4.ps written in PostScript format)

107
108 /*******************************************************/
109 /** Linear Regression                                  **/
110 /*******************************************************/
111
112
113 svyset masterid [pweight=wtfinal], strata(stratum)

        pweight: wtfinal
            VCE: linearized
    Single unit: missing
       Strata 1: stratum
           SU 1: masterid
          FPC 1: <zero>

114 svy: regress sqrtincome hoursworked
  (running regress on estimation sample)

  Survey: Linear regression

  Number of strata   =        90                Number of obs      =     18742
  Number of PSUs     =     18742                Population size    = 3849870.5
                                                 Design df          =     18652
                                                 F(   1,  18652)    =    135.08
                                                 Prob > F           =    0.0000
                                                 R-squared          =    0.0178
```

| sqrtincome | Coef. | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| hoursworked | **1.025276** | **.0882157** | **11.62** | **0.000** | **.8523655** | **1.198187** |
| _cons | **186.6406** | **3.777511** | **49.41** | **0.000** | **179.2364** | **194.0449** |

```
115
116 /*NOT THE SAME AS : no design information*/
117 regress sqrtincome hoursworked [pweight=wtfinal]
  (sum of wgt is   3.8499e+06)

  Linear regression                               Number of obs =   18742
                                                  F(  1, 18740) =  135.25
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.0178
                                                  Root MSE      =  90.918
```

| sqrtincome | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| hoursworked | **1.025276** | **.0881609** | **11.63** | **0.000** | **.852473** | **1.19808** |
| _cons | **186.6406** | **3.776342** | **49.42** | **0.000** | **179.2387** | **194.0426** |

```
118
119 /*NOT THE SAME AS : weighted least squares*/
120 regress sqrtincome hoursworked [aweight=wtfinal]
  (sum of wgt is   3.8499e+06)
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | **2814492.5** | **1** | **2814492.5** | Number of obs = | **18742** |
| Residual | **154905534** | **18740** | **8266.03702** | F( 1, 18740) = | **340.49** |
| | | | | Prob > F = | **0.0000** |
| Total | **157720026** | **18741** | **8415.77431** | R-squared = | **0.0178** |
| | | | | Adj R-squared = | **0.0178** |
| | | | | Root MSE = | **90.918** |

| sqrtincome | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| hoursworked | **1.025276** | **.0555635** | **18.45** | **0.000** | **.9163669** | **1.134186** |
| _cons | **186.6406** | **2.324936** | **80.28** | **0.000** | **182.0835** | **191.1977** |

```
121
122
123
124 /*******************************************************/
125 /** Logistic Regression                                **/
126 /*******************************************************/
127
128 gen uninsured = 0

129 replace uninsured = 1 if a1 == 2
  (5400 real changes made)

130 replace uninsured = . if a1 > 2
  (31 real changes made, 31 to missing)
```

```
131 replace uninsured = . if a1 == .
  (0 real changes made)

132
133 svy: logistic uninsured hoursworked
  (running logistic on estimation sample)

  Survey: Logistic regression

  Number of strata   =        90                  Number of obs      =     23184
  Number of PSUs     =     23184                  Population size    = 4795619.1
                                                  Design df          =     23094
                                                  F(   1,  23094)    =     46.98
                                                  Prob > F           =    0.0000
```

| uninsured | Odds Ratio | Linearized Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| hoursworked | **.9827379** | **.0024965** | **−6.85** | **0.000** | **.9778567** | **.9876434** |

```
134
135
136
137 /*******************************************************/
138 /** Small Area Estimation                             **/
139 /*******************************************************/
140
141 clear

142
143 set memory 1g  /*increase available memory, since the data is huge*/
  (1048576k)

144 use "ofhs_virgin_file.dta"   /*read in the data*/

145
146 /*Example:  consider Minorities in Champaign County */
147 /*race_aim is the self-reported race*/
148 gen minority = 1

149 replace minority = 0 if race_aim == 1
  (33740 real changes made)

150 replace minority = . if race_aim == .
  (0 real changes made)

151
152 /*Make a new variable to identify this "small area"*/
153 /*s9 is self reported county != stratum, 21=Champaign Co.*/
154
155 gen smallarea = 0

156 replace smallarea = 1 if minority == 1 & s9 == 21
  (6 real changes made)

157
158 /*recode the insurance status*/
159 gen insured = a1
```

```
160 replace insured = 0 if insured == 2
  (5400 real changes made)

161 replace insured = . if insured >= 98
  (31 real changes made, 31 to missing)

162 tab insured
```

| insured | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | **5,400** | **13.53** | **13.53** |
| 1 | **34,522** | **86.47** | **100.00** |
| Total | **39,922** | **100.00** | |

```
163
164 gen uninsured = 1 - insured
  (31 missing values generated)

165
166 /*create a counter variable*/
167 gen const = 1

168
169 svyset masterid [pweight=wtfinal], strata(stratum)

        pweight: wtfinal
            VCE: linearized
    Single unit: missing
       Strata 1: stratum
           SU 1: masterid
          FPC 1: <zero>

170
171 /******************/
172 /*Direct Estimator*/
173
174 svy: ratio insuredratiosmallarea: uninsured/const, subpop(smallarea)
  (running ratio on estimation sample)

  Survey: Ratio estimation

  Number of strata =       2          Number of obs    =    1389
  Number of PSUs   =    1389          Population size  =  119526
                                      Subpop. no. obs  =       6
                                      Subpop. size     = 1486.53
                                      Design df        =    1387

   insuredrat~a: uninsured/const
```

| | Ratio | Linearized Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| insuredrat~a | **.1131079** | **.1238076** | **-.1297624** | **.3559782** |

```
  Note: 88 strata omitted because they contain no subpopulation
        members.

175 lincom _b[insuredratiosmallarea] * 1062 /*multiply estimate by the Minorioty Population in Champai

   ( 1)  1062 insuredratiosmallarea = 0
```

| | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| (1) | **120.1206** | **131.4836** | **0.91** | **0.361** | **-137.8077** | **378.0489** |

```
176 scalar t_hat_dir = r(estimate)

177 scalar t_dir_var = r(se)^2

178
179
180 /******************/
181 /*Synthetic Estimator*/
182
183 /*assume proportion of insured is constant across the state, and uniform over minority subpop*/
184
185 svy: ratio insuredratio: uninsured/const
  (running ratio on estimation sample)

  Survey: Ratio estimation

  Number of strata =      90          Number of obs    =   39922
  Number of PSUs   =   39922          Population size  = 8.5e+06
                                      Design df        =   39832

   insuredratio: uninsured/const

  ------------------------------------------------------------------------------
               |                Linearized
               |       Ratio   Std. Err.     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
  insuredratio |   .1388642   .0023893      .1341811    .1435473
  ------------------------------------------------------------------------------

186 lincom _b[insuredratio] * 1062 /*multiply estimate by the Minorioty Population in Champaign County

   ( 1)  1062 insuredratio = 0

  ------------------------------------------------------------------------------
               |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
           (1) |   147.4738   2.537454    58.12   0.000     142.5003    152.4473
  ------------------------------------------------------------------------------

187 scalar t_hat_syn = r(estimate)

188 scalar t_syn_var = r(se)^2

189
190 /*assume proportion of insured is constant in the minority population across the state*/
191
192 svy: ratio insuredratiominority: uninsured/const, subpop(minority)
  (running ratio on estimation sample)

  Survey: Ratio estimation

  Number of strata =      88          Number of obs    =   39712
  Number of PSUs   =   39712          Population size  = 8.5e+06
                                      Subpop. no. obs  =    6205
                                      Subpop. size     = 1.1e+06
                                      Design df        =   39624

   insuredrat~y: uninsured/const

  ------------------------------------------------------------------------------
               |                Linearized
               |       Ratio   Std. Err.     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
  insuredrat~y |   .2300258   .0079734      .2143977    .2456539
  ------------------------------------------------------------------------------
  Note: 2 strata omitted because they contain no subpopulation
        members.
```

```
193 estimates store syntheticest

194
195 lincom _b[insuredratiominority] * 1062 /*multiply estimate by the Minorioty Population in Champaig

   ( 1)  1062 insuredratiominority = 0
```

| | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| (1) | **244.2874** | **8.467789** | **28.85** | **0.000** | **227.6903** | **260.8845** |

```
196 scalar t_hat_syn2 = r(estimate)

197 scalar t_syn2_var = r(se)^2

198
199
200
201 /***********************/
202 /* Composite Estimator */
203
204 scalar alpha = 0.5

205 scalar t_hat_comp = alpha * t_hat_dir + (1-alpha) * t_hat_syn2

206
207
208 /***********************/
209 /* Compare Estimates   */
210
211 display t_hat_dir
  120.12058

212 display t_hat_syn2
  244.28739

213 display t_hat_comp
  182.20398

214
215
216
  end of do-file

217 log close
      log:  /scratch/OhioFamilyHealthSurvey/analysislogModel.smcl
  log type:  smcl
 closed on:  16 Jul 2008, 16:06:06
```