USING STATA TO ANALYZE DATA FROM A SAMPLE SURVEY

Kim Chantala
Carolina Population Center
UNC Chapel Hill

Last Update: October 1, 2001

A tutorial for understanding the basic Stata survey capabilities is available on the CPC web-page: http://www.cpc.unc.edu/services/computer/presentations/statatutorial/. If you are unfamiliar with these capabilities, it will be helpful if you have gone through the tutorial before using this document.

WARNING! Results from the examples are for illustrating usage of software and may not be representative of actual findings. These results should not be quoted.

INTRODUCTION

A sample survey is conducted to obtain information about the characteristics of a population. To reduce the cost and time necessary to collect the data, this task is often handled by selecting a subset (*a sample*) from the set of all measurements (*the target population*) of interest to the researchers. The methods that are used to select the sample often incorporate stratification, clustering and unequal probability of selection of participants. These characteristics must be incorporated into your analysis to obtain unbiased estimates concerning the entire population. This paper demonstrates appropriate analytical techniques from Stata specifically designed to handle these added complexities via a series of examples using data from a sample survey. Techniques covered will include contingency tables, regression, logistic regression, ordered logistic regression, multinomial logistic regression, and time-to-event analysis.

GUIDELINES FOR ANALYZING DATA FROM A SAMPLE SURVEY

To analyze data from a sample survey, you need to make sure that the design features of the sampling strategy are included with your analysis file. You must also account for the effects of sampling design on population estimates by using analytical methods appropriate for handling correlated data collected with unequal probability of selection. Failure to account for the sampling design usually leads to under-estimating standard errors and false-positive statistical test results.

This section presents guidelines adapted from "Sampling of Populations: Methods and Applications" by Paul S. Levy and Stanley Lemeshow, 1999, John Wiley and Sons that will help you get your analysis done correctly.

Getting the data ready for analysis.

Step 1. Identify the variables that describe the sample design. Find the variables that describe the following characteristics:

- Stratification variable. Stratification is the division of a population into mutually exclusive parts (strata) for the purpose of drawing a sample. A proportion of the sample is selected from each stratum. Stratification may be made on a geographic basis or by reference to some other quality of the population.
- *Cluster (Primary Sampling Unit) variable*. This is needed because participants sampled from the same cluster are likely to respond more alike than participants sampled from different clusters.
- *Sampling (Probability) Weight variable*. Recall that:
  Sampling Weight = inverse of the selection probability
  = # of subjects in the population represent by participant.
- *Population Number of PSUs per Stratum variable.* This variable may not be needed if sampling with replacement can be assumed. With Replacement (WR) sampling means that after a sampling unit is selected it is returned to the population before the next unit is

selected. In the contrary case, the sampling is "Without Replacement (WOR)". If the sampling fraction (# selected/# eligible) for each strata is less than 0.1 then the difference between WOR and WR is small and you can treat your design as a WR design.

Step 2. Make sure the variables in step 1 are available on each observation in the data set. If there are observations that have missing values for the design variables check with your data manager to determine why this is happening.

Step 3. Create any analysis variables you need. If you are creating variables that include only information available for each participant, then you do not have to incorporate any of the design variables. However, if you want to use the participants' data to construct cluster-level variables, then you should consider incorporating the sampling weights. For example, if you have test scores for students (the participant) sampled from a school (the primary sampling unit), then to compute the average test score for schools like each one in your sample, you would want to use the sampling weight to compute the weighted average score.

Step 4. Create sub-population variable. Identify the population you are interested in analyzing and create an appropriate indicator variable to use for specifying the sub-population. Recall from the tutorial, that the *svy* commands in Stata need information from the every observation in the data set to correctly compute variance, standard error, confidence intervals, and p-values.

Running Your Analysis.

The next step is to determine the best set of commands for performing the desired analysis. Stata provides two ways to analyze survey data. Appendix A also contains templates to illustrate how to set up your analysis with each method. You also can find out more about each method by going through the Stata tutorial. Here is a table from that tutorial to help you decide which one to choose.

Table 1. A comparison of methods to analyze data from a sample survey.

Method	Strengths	Limitations
The *svy* commands	*svytest* and *svylc* commands used after estimation adjust the test statistics correctly for the sample design.	Cannot subset data - the *subpop* option must be used for sub-population analysis.
	Can make finite population corrections for without replacement samples.	Not all types of analysis available with the *svy* commands.
	Option available on *svyset* command to specify the stratification variable.	
Commands that allow *pweight* and *robust cluster()* options	The data set can be subset for sub-population analysis.	Should have at least 40 clusters available.
	Additional estimation commands that do not have an analogous *svy* command, such as cox regression are available	Option for specifying a stratification variable is not available.
		test and *lincom* do not adjust the test statistics for the sample design. If the number of clusters is large, then this adjustment would be minor.

Here are some common errors that you can avoid.

- Ignoring clustering and unequal probability of selection of participants in your analyses. This results in biased estimates and false-positive hypothesis test results. *Avoid this error by using the svy commands for your analysis. If your analysis technique is not available with the svy commands, then use a command that allows pweight with the robust cluster() option.*
- Using the wrong weight specification in Stata. *For data from a sample survey, you should use the pweight command to define the sampling weight***.** Using any of the other weight commands (aweight, fweight, or iweight) can result in incorrect variance, standard errors, confidence intervals, and p-values.
- Subsetting the sample when using the svy commands in stata. These commands use the Taylor Series approximation for the variance estimation and must be able to correctly count the number of primary sampling units (PSUs) that were originally sampled. Subsetting the data may cause an incorrect number of PSU's to be used in the variance computation formula**.** *Do not subset the data from a sample survey and always use the subpop option when using the svy commands to do sub-population analysis*.

DATA USED IN EXAMPLES

The following examples utilize data from the National Longitudinal Survey of Adolescent Health (Add Health). Add Health is a panel study of adolescents selected from enrollment rosters for grades 7 through 12 from a sample of schools in the United States. Both schools (primary sampling unit) and adolescents were selected with unequal probabilities. Add Health data can be analyzed as having been selected by using a "with replacement" sampling strategy. Interviews were in 1994 (In-school), 1995 (Wave I In-home), 1996 (Wave II In-home) and 2001 (Wave III In-home).

The following variables used in the examples were constructed from the Wave I In-home questionnaire of the National Longitudinal Survey of Adolescent Health (Add Health).

Table 2. Variables used in examples.

Use	Variable Name	Meaning
Design Variables	REGION	Strata Variable – Region of country 1=West 3=South 2=Midwest 4=Northeast
	GSWGT1	Sample Weight from Wave I
	PSUSCID	Primary Sampling Unit (PSU) – School ID
Covariates	MALE	Male Adolescent? 0=No, Adolescent is female 1=Yes, Adolescent is male
	AGE_KID	Integer Age of Respondent: 11-21 years old
	BIOMAPA	Live with Bio Mom & Dad 0=No 1=Yes
	ENGL_GPA	English Grade A=4, B=3, C=2, D/F=1
	BLACK_	Black race? 0=No , 1=Yes, Black only or Black + other races
	WHITE_	White race? 0=No , 1=Yes, White only or White + other races
	HISPANIC	Hispanic ethnic group? 0=No , 1=Yes
	SMKCIG	Smokes Cigarettes regularly? 0=No , 1=Yes
	BACKGRAD	Ever held back a grade? 0=No, 1=Yes
	TRBTEACH	Have you had trouble getting along with teachers? 0=Never, 1=A few times, 2=Almost once a week 3=Almost everyday, 4=Everyday
Outcome Variables	PVT_PER1	Age Standardized Percentile Rank for Add Health Picture Vocabulary Test . Range of Values: 1-100
	PVT_Q4	Age Standardized Percentile Rank for Add Health Picture Vocabulary Test is 75^{th} percentile or greater 0=under 75^{th} percentile 1=75^{th} percentile or greater
	PVTQ1_4	Age Standardized Percentile Rank Quartile 1= under 25^{th} percentile, 2=25^{th} to 49^{th} percentile 3=50^{th} to 74^{th} percentile, 4=75^{th} percentile & over

REGRESSION ANALYSIS

This example illustrates the use of commands from Stata that can be used to perform a multiple regression analysis. A partial listing of the output is included with an interpretation of results.

Research Question:

Is performance on the Add Health Vocabulary test influenced by an adolescent's sex or family composition?

Predictive Model:

$$PVT\_PER1 = b_0 + b_1 AGE\_KID + b_2 MALE + b_3 BIOMAPA$$

Where

b_0 = Intercept
b_1 = Change in Test score for one year increment in age
b_2 =Difference in Test Score between males and females
b_3 =Change in Test Score for living with Biological Parents

The following table shows the model for each level of the categorical variables:

MALE	BIOMAPA	Prediction Equation
0=No	0=No	$PVT\_PER1 = b_0 + b_1 * AGE\_KID + b_2 * 0 + b_3 * 0 = b_0 + b_1 * AGE\_KID$
1=Yes	0=No	$PVT\_PER1 = b_0 + b_1 * AGE\_KID + b_2 * 1 + b_3 * 0 = (b_0 + b_2) + b_1 * AGE\_KID$
0=No	1=Yes	$PVT\_PER1 = b_0 + b_1 * AGE\_KID + b_2 * 0 + b_3 * 1 = (b_0 + b_3) + b_1 * AGE\_KID$
1=Yes	1=Yes	$PVT\_PER1 = b_0 + b_1 * AGE\_KID + b_2 * 1 + b_3 * 1 (b_0 + b_2 + b_3) + b_1 * AGE\_KID$

The equation for each group defined by sex and living with biological parents has a different intercept, but a common slope for age of adolescent.

Using Stata: SVYREG

STATA CODE

The *svyset* command is used to specify the design information for analysis. Use the *strata* keyword to specify the stratification variable (*region)*, the *pweight* keyword to specify the weight variable (*gswgt1)*, and the *psu* keyword to specify the primary sampling unit (*psuscid).*

The first variable following the *svyreg* command denotes the outcome of our model and the following variables are the covariates.

```
svyset strata region
svyset pweight gswgt1
svyset psu psuscid
svyreg pvt_per1 age_kid male biomapa
```

STATA OUTPUT

The *svyset* command uses Taylor series linearization methods to estimate variance and standard error of the estimates. The R-squared statistic indicates only 3% of the variation in the data is explained by the model. Age of adolescent is not associated with Percentile PVT Score, but males are predicted to have 2.18 percentage points higher scores than females, while adolescents living with both biological parents will have a score 8.96 points higher than adolescent's not living with their biological parents.

```
Survey linear regression

pweight:  gswgt1                                  Number of obs    =     18001
Strata:   region                                  Number of strata =         4
PSU:      psuscid                                 Number of PSUs   =       132
                                                  Population size  =  21095781
                                                  F(   3,    126)  =     38.08
                                                  Prob > F         =    0.0000
                                                  R-squared        =    0.0304

------------------------------------------------------------------------------
pvt_per1 |      Coef.    Std. Err.       t     P>|t|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
 age_kid |  -.5964347     .469777     -1.270   0.207     -1.525969    .3330993
    male |   2.178519    .6033584      3.611   0.000      .9846715    3.372367
 biomapa |    8.95549    .9117622      9.822   0.000      7.151413    10.75957
   _cons |   54.09994    7.220088      7.493   0.000      39.81376    68.38612
------------------------------------------------------------------------------
```

Understanding the Predictive Model

The following table shows the prediction equation for each level of the qualitative variables.

MALE	BIOMAPA	Prediction Equation
0=No	0=No	$PVT\_PER1 = b_0 + b_1 * AGE\_KID + b_2 * 0 + b_3 * 0$ $= 54.10 - 0.596 * AGE\_KID$
1=Yes	0=No	$PVT\_PER1 = b_0 + b_1 * AGE\_KID + b_2 * 1 + b_3 * 0$ $= 54.10 - 0.596 * AGE\_KID + 2.179$ $= 56.279 - 0.596 * AGE\_KID$
0=No	1=Yes	$PVT\_PER1 = b_0 + b_1 * AGE\_KID + b_2 * 0 + b_3 * 1$ $= 54.10 - 0.596 * AGE\_KID + 8.955$ $= 63.055 - 0.596 * AGE\_KID$
1=Yes	1=Yes	$PVT\_PER1 = b_0 + b_1 * AGE\_KID + b_2 * 1 + b_3 * 1$ $= 54.10 - 0.596 * AGE\_KID + 2.179 + 8.955$ $= 65.23 - 0.596 * AGE\_KID$

To estimate the difference between males who live with their biological parents and females who do not live with their biological parents, you just subtract the two prediction equations:

$$Difference = PVT\_PER1_{Male=1,BIOMAPA=1} - PVT\_PER1_{Male=0,BIOMAPA=0}$$
$$= (b_0 + b_1 * AGE\_KID + b_2 + b_3) - (b_0 + b_1 * AGE\_KID)$$
$$= b_2 + b_3 = 2.179 + 8.955 = 11.13$$

Males who live with both biological parents score 11.13 points higher than girls who do not. You can use the *svylc* (survey linear combination) or *svytest* (survey test) command in Stata to estimate this linear combination ($H_0: b_2 + b_3 = 0$) and compute a p-value for the testing the linear combination is zero.

CONTINGENCY TABLES

Often the response of interest in the Add Health data is categorical in nature. This section illustrates how to present categorical data in a contingency table and strategies for assessing association using Stata.

Research Question:

Is there an association between living with your biological parents and the quartile for the score received on the Add Health Picture Vocabulary Test?

To answer this question, we will look for an association between the variables *biomapa* and *pvtq1_4*.

Using Stata: svytab

STATA CODE

The *svyset* command is used to define the stratification variable (*region*), sampling weight (*gswgt1*) and the primary sampling unit (*psuscid*). The *svytab* command will create a contingency table showing the values of *biomapa* down the side and the values of *pvtq1_4* across the top. Options specified here for *svytab* request that the table display cell proportions (cell), standard errors of cell proportions (*se*), and 95% confidence intervals for cell proportions (*ci*).

```
svyset strata region
svyset pweight gswgt1
svyset psu psuscid
svytab biomapa pvtq1_4, cell se ci
```

STATA OUTPUT

Stata reports that 18,001 respondents representing 21,095,781 adolescents have complete data and are used in the computations for the contingency table. Each cell in the table contains three statistics:

- Population estimates for the proportion of adolescents scoring in a particular PVT quartile and presence (or absence) of both biological parents in the home
- the standard error of the proportion

- the 95% confidence interval of the proportion.

```
pweight:  gswgt1                                  Number of obs      =     18001
Strata:   region                                  Number of strata   =         4
PSU:      psuscid                                 Number of PSUs     =       132
                                                  Population size    =  21095781

------------------------------------------------------------------------------
Live     |
with     |
Bio Mom  |
& Dad    |                              PVT quartile
0=N/1=Y  |             1             2             3             4         Total
---------+--------------------------------------------------------------------
   0=No  |         .1149         .1542         .1157         .0822         .4671
         |       (.0106)       (.0066)       (.0043)       (.0044)       (.0129)
         | [.0955,.1375] [.1417,.1677] [.1074,.1246] [.0739,.0914] [.4417,.4926]
         |
  1=Yes  |         .0799         .1477         .1632         .1421         .5329
         |       (.0066)       (.0061)       (.0072)       (.0097)       (.0129)
         | [.0678,.0939]  [.136,.1602]  [.1494,.178] [.1239,.1624] [.5074,.5583]
         |
  Total  |         .1948         .3019          .279         .2243             1
         |       (.0148)       (.0084)       (.0089)       (.0129)
         | [.1672,.2257] [.2857,.3187] [.2618,.2969] [.1998,.2509]
------------------------------------------------------------------------------

  Key:  cell proportions
        (standard errors of cell proportions)
        [95% confidence intervals for cell proportions]

  Pearson:
    Uncorrected   chi2(3)         =  472.5874
    Design-based  F(2.64, 338.04) =   48.4521     P = 0.0000
```

The summary statistics (Pearson) show the uncorrected chi-square and the corrected design-based F-statistic to evaluate the association between living with both biological parents and PVT test score quartile. This F-statistic shows that the association is significant ($p \leq 0.0001$).

Stata has many options to control the contents of the table. For example, the stata command:

```
svytab biomapa pvtq1_4, obs row count
```

can be used to obtain population estimates of the number of adolescents in each cell (weighted count) the number of respondents (observations) in each cell, and the row proportions.

LOGISTIC REGRESSION ANALYSIS

Logistic Regression is used to model dichotomous (0 or 1) outcomes. This technique models the log odds of an outcome defined by the values of covariates in your model. In addition to

covering how to model sub-populations, we will use both the svy commands and the robust cluster commands.

Research Question:

How is being in the upper quartile of the Vocabulary test score influenced by a boy's grade in English and Family composition?

Predictive Model:

$$\log\left(\frac{\Pr(PVT\_Q4=1)}{1-\Pr(PVT\_Q4=1)}\right) = b_0 + b_1 AGE\_KID + b_2 BIOMAPA + b_3 ENGL\_GPA$$

Where

b_0 = Intercept

b_1 = Change in log odds of being in upper quartile for one year increment in age

b_2 =Change in log odds of being in upper quartile for living with Biological Parents

b_3 =Change in log odds of being in upper quartile for increase in one grade level

The model predicted log-odds for the categorical subpopulations will be:

BIOMAPA	ENGL_GPA	Ln(odds)
0=No	4=A	$\beta_0 + \beta_1$ AGE_KID + $4\beta_3$
0=No	3=B	$\beta_0 + \beta_1$ AGE_KID + $3\beta_3$
0=No	2=C	$\beta_0 + \beta_1 *$AGE_KID + $2\beta_3$
0=No	1=D/F	$\beta_0 + \beta_1 *$AGE_KID + β_3
1=Yes	4=A	$\beta_0 + \beta_1$ AGE_KID +β_2 + $4\beta_3$
1=Yes	3=B	$\beta_0 + \beta_1$ AGE_KID +β_2 + $3\beta_3$
1=Yes	2=C	$\beta_0 + \beta_1 *$AGE_KID +β_2 + $2\beta_3$
1=Yes	1=D/F	$\beta_0 + \beta_1 *$AGE_KID +β_2 + β_3

We are assuming a model with a common slope for age of the boy, but different intercepts defined by grade in English and living with both biological parents.

The relationship between probability and odds

The odds of an outcome is related to the probability of the outcome by the following relation:

$$odds = \frac{probability}{1 - probability}$$

An odds ratio is just the ratio of the odds of the outcome evaluated at two different sets of values for your covariates. It is easy to show that to test the hypothesis that $p_1 = p_2$ you can test that the hypothesis that an odds ratio comparing group 1 to group 2 is equal to 1.

However, you cannot easily put a confidence interval on the difference between the two probabilities.

Using Stata: SVYLOGIT

STATA CODE

The *svyset* command is used to specify the design information for analysis. Use the *strata* keyword to specify the stratification variable, *region,* the *pweight* keyword to specify the weight variable *gswgt1*, and the *psu* keyword to specify the primary sampling unit, *psuscid.* The first variable following the *svylogit* command denotes the outcome (*pvt_q4*) of our model and the following variables are the covariates. The option *subpop* is used to specify the sub-population we want to be used to compute parameter estimates. All 18,924 observations are needed for the variance computation because Stata determines the design information (number of primary sampling units) used in the formula variance computation

```
svyset strata region
svyset pweight gswgt1
svyset psu psuscid
svylogit pvt_q4 age_kid biomapa engl_gpa, subpop(male)
```

STATA OUTPUT

Stata lists the number observations with no missing values for the variables in the model (N=17,191) and has summed the corresponding sample weights to estimate 19,955,620 adolescents in the U.S. are represented by these observations. The number of observations with complete data in the sub-population is 8,366 representing 10,084,117 boys. Note that the number of strata (4) and primary sampling units (132) has been correctly counted.

```
Survey logistic regression
pweight:  gswgt1                                  Number of obs    =     17191
Strata:   region                                  Number of strata =         4
PSU:      psuscid                                 Number of PSUs   =       132
                                                  Population size  =  19955620
Subpopulation no. of obs =      8366              F(   3,    126)  =     49.14
Subpopulation size       =  10084117              Prob > F         =    0.0000
------------------------------------------------------------------------------
  pvt_q4 |      Coef.    Std. Err.       t     P>|t|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
 age_kid |  -.0451845    .0278879     -1.620   0.108     -.1003656    .0099965
 biomapa |   .4273138    .0820139      5.210   0.000      .2650354    .5895923
engl_gpa |   .4258579    .0423055     10.066   0.000      .3421493    .5095664
   _cons |  -1.886177    .4411884     -4.275   0.000     -2.759144   -1.013211
```

The adjust command can be used to estimate a linear combination of the coefficients estimated for the variables in our model. If you do not specify a value for a variable when using *adjust*, Stata will incorrectly substitute the sample mean rather than an estimate of the population mean. This is because *adjust* ignores any weights used by the estimation commands. To correctly compute a linear combination, it is necessary to specify a value for all variables in the model. For example, the following statement:

```
adjust age_kid=17 engl_gpa=3, by(biomapa) xb se ci
```

produces an estimate of the log odds of scoring above the 75th percentile for boys at age 17 with a grade of B in English for both categories of living with both biological parents:

```
------------------------------------------------------------------------------
     Dependent variable: pvt_q4     Command: svylogit
Covariates set to value: age_kid = 17, engl_gpa = 3
------------------------------------------------------------------------------

-------------------------------------------------------
Live with |
Bio Mom & |
Dad       |
0=N/1=Y   |         xb        stdp          lb          ub
----------+--------------------------------------------
        0 |   -1.37674   (.100564)   [-1.57572   -1.17776]
        1 |   -.949427   (.094665)   [-1.13674   -.762115]
-------------------------------------------------------
     Key:  xb         =  Linear Prediction
           stdp       =  Standard Error
           [lb , ub]  =  [95% Confidence Interval]
```

You can also include the *exp* option at the end of the adjust command to get adjust to print exponentiated linear combination of the coefficients. The *pr* option on *adjust* is not available after using the *svylogit* command.

The svylc command can also be used to produce linear combinations of the coefficients:

```
. svylc 17*age_kid + 1*biomapa + 3*engl_gpa + _cons

 ( 1)  17.0 age_kid + biomapa + 3.0 engl_gpa + _cons = 0.0

------------------------------------------------------------------------------
      pvt_q4 |      Coef.    Std. Err.      t    P>|t|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.9494267    .0946653   -10.03   0.000   -1.136738   -.7621154
------------------------------------------------------------------------------
```

The results from svylc match those from adjust. The advantage of using svylc is that a hypothesis test can also be performed. For example, suppose you want to compute the odds ratio comparing 17 year-old boys not living with both biological parents to12 year-old boys living with both biological parents. Assume both boys make the same grade in English. We would want to estimate the difference in log odds for these to :

$$(\beta_0 + 17*\beta_1 + \text{GRADE}*\beta_3) - (\beta_0 + 12*\beta_1 + \beta_2 + \text{GRADE}*\beta_3) = 5*\beta_1 - \beta_2$$

Since β_1 is the coefficient for AGE_KID and β_2 is the coefficient for BIOMAPA, the *svylc* command would be:

```
. svylc 5*age_kid - 1*biomapa
```

This produces the desired difference in log odds:

```
 ( 1)  5.0 age_kid - biomapa = 0.0

------------------------------------------------------------------------------
      pvt_q4 |      Coef.    Std. Err.      t    P>|t|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.6532364    .1641137    -3.98   0.000   -.9779635   -.3285094
------------------------------------------------------------------------------
```

The *or* option can be added to the svylc command to get the odds ratio ($e^{5*\beta 1-\beta 2}$):

```
. svylc 5*age_kid - 1*biomapa , or
```

The following table will be printed.

```
 ( 1)  5.0 age_kid - biomapa = 0.0

------------------------------------------------------------------------------
      pvt_q4 | Odds Ratio    Std. Err.      t    P>|t|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .5203589     .085398    -3.98   0.000    .3760762    .7199962
------------------------------------------------------------------------------
```

Thus, assuming equal grades in English, the odds of a 17 year-old boy not living with both biological parents is only half that of a 12 year boy who lives with his biological parents.

The svytest command can be used to test joint hypothesis about variables. For example, testing that the coefficient for age_kid and biomapa are both equal to zero can be done with the following stata command:

```
. svytest age_kid biomapa
```

which produces the following output:

```
Adjusted Wald test

 ( 1)  age_kid = 0.0
 ( 2)  biomapa = 0.0

       F(  2,   127) =   14.51
            Prob > F =    0.0000
```

Using STATA: logit with pweight and robust cluster

STATA CODE

Note that we can subset the data *(if male==1)* when using the *robust cluster( )* options in stata and still have the variance computed with an acceptable technique. The primary sampling unit *(psuscid)* is used as the argument to the *cluster* option and the sample weights *(gswgt1)* are specified by *[pweight=gswgt1]*.

```
logit pvt_q4 age_kid biomapa engl_gpa if male==1 [pweight=gswgt1],
robust cluster(psuscid)
```

STATA OUTPUT

The results and interpretation in the following output are identical to the results from svylogit.

```
Logit estimates                                   Number of obs   =       8366
                                                  Wald chi2(3)    =     142.44
                                                  Prob > chi2     =     0.0000
Log likelihood = -4429.7883                       Pseudo R2       =     0.0384

                          (standard errors adjusted for clustering on psuscid)
------------------------------------------------------------------------------
             |               Robust
      pvt_q4 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_kid |  -.0451845   .0277835     -1.626   0.104      -.0996393    .0092702
     biomapa |   .4273138   .0817512      5.227   0.000       .2670844    .5875433
    engl_gpa |   .4258579   .0430438      9.894   0.000       .3414936    .5102222
       _cons |  -1.886177   .4414855     -4.272   0.000      -2.751473   -1.020881
------------------------------------------------------------------------------
```

Linear combinations of parameter estimates can be obtained with the *lincom* command and hypothesis tests can be performed with the *test* command. The results would be the same as those obtained with *svylc* and *svytest*. Adjust can also be used in the same manner as with svylogit.

ORDERED LOGISTIC REGRESSION ANALYSIS

Ordered logistic regression models can be used to predict the relationship between multi-level ordinal outcomes and a set of explanatory variables. For example, we may wish to predict the quartile of achievement (under 25^{th}, 25^{th} to 49^{th}, 50^{th} to 74^{th}, and 75^{th} percentile & over) on the Add Health Vocabulary test. We could collapse these categories to create three binary outcomes:

Pr{under 25^{th} percentile} vs. Pr{over 25^{th} percentile}

Pr{under 50^{th} percentile} vs. Pr{over 50^{th} percentile}

Pr{under 75^{th} percentile} vs. Pr{over 75^{th} percentile}

Standard logistic regression models could then be used to fit a separate model to each of these outcomes. However, it makes sense to take advantage of the natural ordering of these categories by estimating the three models simultaneously using proportional odds models. These models can be estimated with ordered logistic regression analysis. The basic assumption made to do this type of analysis is that the regression lines for the different outcomes are parallel to each other but are allowed to have different intercepts.

These models can be fit using either SUDAAN or Stata. However, SUDAAN and Stata use a different model for estimating the parameters. The models estimated by SUDAAN have the form:

$$\log\left(\frac{\Pr(category\ j\ or\ lower)}{\Pr(category\ j+1\ or\ higher)}\right)=k_j+\sum_{i=1}^{p}b_i x_i$$

This is the same formula that SAS uses. The models estimated by Stata have the form:

$$\log\left(\frac{\Pr(category\ j\ or\ lower)}{\Pr(category\ j+1\ or\ higher)}\right)=k_j-\sum_{i=1}^{p}b_i x_i$$

For both models, the κ_j's are the intercepts or "cut-points" and will be estimated to have the same value by both SUDAAN or Stata. The β_i's in the summation are regression coefficients for the covariates x_i. Notice that the SUDAAN or SAS model adds the summation while Stata subracts the summation. Thus, Stata estimates the β_i's to be equal in magnitude, but opposite in sign than those estimated by SUDAAN or SAS. This means that the odds ratios estimated from the β_i's computed with Stata will be the reciprocal of the odds ratios estimated from the SUDAAN or SAS results. Hence, the interpretation will be different because the reference and the risk groups will be interchanged. For example, if a variable smoking is coded as 0=Non-smoker and 1=Smokers, then the exponentiated coefficient for smoker estimated by Stata will be the odds ratio for Non-smokers to Smokers. When estimated by SUDAAN or SAS, the exponentiated coefficient will represent the odds ratio for Smokers to Non-smokers. Note that in most other estimation commands use the same model as SUDAAN and SAS.

Research Question:

How is the quartile of achievement in the Vocabulary test influenced by a boy's grade in English and family composition?

Predictive Models for SUDAAN (or SAS):

$$\log\left(\frac{\Pr(PVT1\_Q4=1)}{\Pr(PVT1\_Q4=2,3,4)}\right)=k_1+b_1 AGE\_KID+b_2 BIOMAPA+b_3 ENGL\_GPA$$

$$\log\left(\frac{\Pr(PVT1\_Q4=1or\,2)}{\Pr(PVT1\_Q4=3or\,4)}\right)=k_2+b_1 AGE\_KID+b_2 BIOMAPA+b_3 ENGL\_GPA$$

$$\log\left(\frac{\Pr(PVT1\_Q4=1or\,2\ or\,3)}{\Pr(PVT1\_Q4=4)}\right)=k_3+b_1 AGE\_KID+b_2 BIOMAPA+b_3 ENGL\_GPA$$

Predictive Models for STATA:

$$\log\left(\frac{\Pr(PVT1\_Q4=1)}{\Pr(PVT1\_Q4=2,3,4)}\right)=k_1-b_1 AGE\_KID-b_2 BIOMAPA-b_3 ENGL\_GPA$$

$$\log\left(\frac{\Pr(PVT1\_Q4 = 1 or\ 2)}{\Pr(PVT1\_Q4 = 3 or\ 4)}\right) = k_2 - b_1 AGE\_KID - b_2 BIOMAPA - b_3 ENGL\_GPA$$

$$\log\left(\frac{\Pr(PVT1\_Q4 = 1 or\ 2\ or 3)}{\Pr(PVT1\_Q4 = 4)}\right) = k_3 - b_1 AGE\_KID - b_2 BIOMAPA - b_3 ENGL\_GPA$$

Interpretation of parameters

Parameter	SUDAAN (or SAS)	Stata
κ_1	Intercept for Log odds of scoring below 25th vs. Over 25th percent	Intercept for Log odds of scoring below 25th vs. Over 25th percent
κ_2	Intercept for Log odds of scoring below 50th vs. Over 50th percent	Intercept for Log odds of scoring below 50th vs. Over 50th percent
κ_3	Intercept for Log odds of scoring below 75th vs. Over 75th percent	Intercept for Log odds of scoring below 75th vs. Over 75th percent
b_1	Change in log odds of being under the percentile cut-point for one year *increase* in age	Change in log odds of being under the percentile cut-point for one year *decrease* in age
b_2	Change in log odds of being under the percentile cut-point for *living with* both biological parents	Change in log odds of being under the percentile cut-point for *not living with* both biological parents
b_3	Change in log odds of being under the percentile cut-point for *increase* in one grade in English	Change in log odds of being under the percentile cut-point for *decrease* in one grade in English

Each main effect in a model is equal to the difference between the model evaluated at two different sets of values of the explanatory variables. The meaning of b_2 in the previous table will be verified as part of this example.

Using Stata: SVYOLOG

STATA CODE

The *svyset* command is used to specify the design information for analysis. Use the *strata* keyword to specify the stratification variable, *region,* the *pweight* keyword to specify the weight variable *gswgt1*, and the *psu* keyword to specify the primary sampling unit, *psuscid.* The first variable following the *svyolog* command denotes the outcome (*pvtq1_4*) of our model and the following variables are the covariates. *Biomapa* is a categorical variables coded as 0=No, 1=Yes, while *age_kid* and *engl_gpa* are treated as continuous variables. The option *subpop* is used to specify the sub-population we want to be used to compute parameter estimates. All 18,924 observations are needed for the variance computation because Stata determines the design information (i.e., number of primary sampling units) used in the formula variance computation.

```
svyset strata region
svyset pweight gswgt1
svyset psu psuscid
svyolog pvtq1_4 age_kid biomapa engl_gpa , subpop(male)
```

STATA OUTPUT

Stata lists the number observations with no missing values for the variables in the model (N=17,191) and has summed the corresponding sample weights to estimate 19,955,620 adolescents in the U.S. The number of observations with complete data in the sub-population is 8,366 representing 10,084,117 boys. Note that the number of strata (4) and primary sampling units (132) has been correctly counted.

```
. svyolog pvtq1_4 age_kid biomapa engl_gpa , subpop(male)

Survey ordered logistic regression
pweight:  gswgt1                                  Number of obs    =     17191
Strata:   region                                  Number of strata =         4
PSU:      psuscid                                 Number of PSUs   =       132
                                                  Population size  =  19955620
Subpopulation no. of obs =      8366              F(   3,    126)  =     70.96
Subpopulation size       =  10084117              Prob > F         =    0.0000
```

The results of the model are listed below.

```
------------------------------------------------------------------------------
     pvtq1_4 |      Coef.    Std. Err.      t    P>|t|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
     age_kid |   -.023944    .0267947    -0.89   0.373   -.0769619    .0290739
     biomapa |   .4953606    .0698794     7.09   0.000    .3570922    .6336289
    engl_gpa |    .321382    .0301858    10.65   0.000    .2616542    .3811098
-------------+----------------------------------------------------------------
       /cut1 |  -.9126946    .4200178    -2.17   0.032   -1.743772   -.0816176
       /cut2 |    .586421    .4253781     1.38   0.170   -.2552622    1.428104
       /cut3 |   1.953444    .4238309     4.61   0.000    1.114823    2.792066
------------------------------------------------------------------------------
```

We can estimate the difference in log odds of scoring lower on the Vocabulary test for *not* living with both biological parents vs. living with both to verify the meaning of b_2 as estimated by Stata:

$$\log\left\{\frac{\Pr(PVT1\_Q4 \le j)}{\Pr(PVT1\_Q4 > j)} \mid BIOMAPA = 0\right\} - \log\left\{\frac{\Pr(PVT1\_Q4 \le j)}{\Pr(PVT1\_Q4 > j)} \mid BIOMAPA = 1\right\}$$

$$= \left\{k_j - b_1 * AGE\_KID - b_2 * 0 - b_3 * ENGL\_GPA\right\} - \left\{k_j - b_1 * AGE\_KID - b_2 * 1 - b_3 * ENGL\_GPA\right\}$$

$$= -(-b_2)$$

$$= b_2$$

Thus, boys not living with both biological parents have $e^{\beta 2} = e^{0.495} = 1.64$ times higher odds of scoring lower than boys living with both biological parents. Similarly, those boys scoring one grade lower in English have $e^{0.321} = 1.38$ times higher odds of scoring lower on the Vocabulary test.

Using SUDAAN

SUDAAN code

The SUDAAN code and an explanation are included here for clarity. The With Replacement design type *(design=WR)* is specified on the *proc* statement. Note that the stratum variable *region* and primary sampling unit variable *psuscid* are listed on the *nest* statement, while the sample weight variable *gswgt1* is listed on the *weight* statement.

SUDAAN requires that all categorical covariates have values 1 or greater. Variable *biomapa* has been recoded as *biomapar* (1=Yes, 2=No). You could have also used the SUDAAN recode statement to recode variables within the procedure. The categorical variables *biomapar* and *pvtq1_4* are listed on the *subgroup* statement while the number of categories for each are listed on the *levels* statement. The *cumlogit* keyword on the model statement fits the proportional odds model.

```
proc multilog data=test filetype=SAS design=WR
     semethod=binder r=independent;
nest region psu;
weight gswgt1;
reflevel  biomapar=2;
subpopn male=1;
subgroup biomapar pvtq1_4;
levels   2       4      ;
model pvtq1_4=age_kid biomapar engl_gpa/cumlogit;
run;
```

SUDAAN results

Note the intercepts computed by SUDAAN have the same value as the intercepts computed by Stata, while the parameter estimates differ by a factor of negative one.

```
Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent             Link Function: Cumulative Logit
Response variable PVTQ1_4: PVT quartile       For Subpopulation: MALE = 1
----------------------------------------------------------------------
PVTQ1_4 (cum-logit),
  Independent                                                 P-value
  Variables and          Beta                                 T-Test
  Effects                Coeff.        SE Beta   T-Test B=0   B=0
----------------------------------------------------------------------
PVTQ1_4 (cum-logit)
  Intercept 1              -0.91         0.42        -2.17     0.0316
  Intercept 2               0.59         0.43         1.38     0.1704
  Intercept 3               1.95         0.42         4.61     0.0000
AGE_KID                     0.02         0.03         0.89     0.3732
Live with bio-parents? 1=Y/2=N
  1                        -0.50         0.07        -7.09     0.0000
  2                         0.00         0.00          .        .
English GPA:
  A=4,B=3,C=2,D/F=1        -0.32         0.03       -10.65     0.0000
```

To understand the meaning of b_2 as estimated by SUDAAN, the difference in log odds of scoring lower for living with both vs. not living with both biological parents gives:

$$\log\left\{\frac{\Pr(PVT1\_Q4 \le j)}{\Pr(PVT1\_Q4 > j)} \mid BIOMAPA = 1\right\} - \log\left\{\frac{\Pr(PVT1\_Q4 \le j)}{\Pr(PVT1\_Q4 > j)} \mid BIOMAPA = 0\right\}$$

$$= \left\{k_j + b_1 * AGE\_KID + b_2 * 1 + b_3 * ENGL\_GPA\right\}$$
$$\quad - \left\{k_j + b_1 * AGE\_KID + b_2 * 0 + b_3 * ENGL\_GPA\right\}$$
$$= b_2$$

Thus, boys living with both biological parents have $e^{\beta 2} = e^{-0.50} = 0.61$ times lower odds of scoring lower than boys not living with both biological parents. Similarly, those boys scoring one grade lower in English have $e^{-0.32} = 0.73$ times higher odds of scoring lower on the Vocabulary test.

MULTINOMIAL REGRESSION ANALYSIS

You may want to model an outcome with multiple levels. This example will show how to fit models to nominal outcomes. Nominal means that there is no ordering to the levels of the outcome. Although the outcome variable used in this is example, PVTQ1_4, is ordered since it is the quartile in which an adolescents PVT test score falls, we will model it as if there is no ordering. The technique selects one level of the outcome as a base category, and models the log of the ratio of being the probability of being in the n^{th} category relative to the base category. This ratio is called the relative risk or odds and the log of this ratio is called the generalized logit.

Research Question:

How is the quartile of the Vocabulary test score for boys influenced by his grade in English and Family composition?

Predictive Model:

$$\log\left(\frac{\Pr(PVTQ1\_4 = 1)}{\Pr(PVTQ1\_4 = 4)}\right) = b_{0,1} + b_{1,1}AGE\_KID + b_{2,1}BIOMAPA + b_{3,1}ENGL\_GPA$$

$$\log\left(\frac{\Pr(PVTQ1\_4 = 2)}{\Pr(PVTQ1\_4 = 4)}\right) = b_{0,2} + b_{1,2}AGE\_KID + b_{2,2}BIOMAPA + b_{3,2}ENGL\_GPA$$

$$\log\left(\frac{\Pr(PVTQ1\_4 = 3)}{\Pr(PVTQ1\_4 = 4)}\right) = b_{0,3} + b_{1,3}AGE\_KID + b_{2,3}BIOMAPA + b_{3,3}ENGL\_GPA$$

Where

$b_{0,n}$ = Intercept for the n^{th} category relative to the 4^{th} (base) category

$b_{1,n}$ = Change in log risk ratio of being in upper quartile for one year increment in age for the n^{th} category relative to the 4^{th} (base) category

$b_{2,n}$ =Change in log risk ratio of being in upper quartile for living with Biological Parents for the n^{th} category relative to the 4^{th} (base) category

$b_{3,n}$ =Change in log risk ratio of being in upper quartile for increase in one grade level for the n^{th} category relative to the 4^{th} (base) category

Using Stata: svymlogit

STATA CODE

The *svyset* command is used to specify the design information for analysis. Use the *strata* keyword to specify the stratification variable, *region,* the *pweight* keyword to specify the weight variable *gswgt1*, and the *psu* keyword to specify the primary sampling unit, *psuscid.*

The first variable following the *svymlog* command denotes the outcome of our model and the following variables are the covariates. The *basecategory* option is used to specify boys in the $75^{th}$and above percentile as the base comparison group.

```
svyset strata region
svyset pweight gswgt1
svyset psu psuscid
svymlog pvtq1_4 age_kid biomapa engl_gpa , subpop(male)
basecategory(4)
```

STATA OUTPUT

Stata lists the number observations with no missing values for the variables in the model (N=17,191) and has summed the corresponding sample weights to estimate 19,955,620 adolescents is the U.S. The number of observations with complete data in the sub-population is 8,366 representing 10,084,117 boys. Note that the number of strata (4) and primary sampling units (132) have been correctly counted.

```
Survey multinomial logistic regression

pweight:  gswgt1                                  Number of obs    =     17191
Strata:   region                                  Number of strata =         4
PSU:      psuscid                                 Number of PSUs   =       132
                                                  Population size  =  19955620
Subpopulation no. of obs =      8366              F(   9,    120)  =     25.25
Subpopulation size       =  10084117              Prob > F         =    0.0000
```

The parameter estimates for each of the three models we fit are given in the output that follows.

```
------------------------------------------------------------------------------
 pvtq1_4 |      Coef.    Std. Err.       t     P>|t|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
1        |
 age_kid |   .0879976    .0487072      1.807   0.073      -.008378    .1843731
 biomapa |  -.7430414    .1251412     -5.938   0.000     -.9906546   -.4954282
engl_gpa |  -.5124962    .0547626     -9.359   0.000     -.6208533    -.404139
   _cons |   .0747199    .7723085      0.097   0.923     -1.453424    1.602864
---------+--------------------------------------------------------------------
2        |
 age_kid |  -.0028244     .030068     -0.094   0.925     -.0623192    .0566703
 biomapa |  -.5457032    .1017309     -5.364   0.000     -.7469952   -.3444112
engl_gpa |  -.4859729     .047102    -10.317   0.000     -.5791724   -.3927735
   _cons |   1.888848    .4763762      3.965   0.000      .9462563     2.83144
---------+--------------------------------------------------------------------
3        |
 age_kid |   .0671961    .0290782      2.311   0.022      .0096598    .1247323
 biomapa |  -.1438588    .0852997     -1.687   0.094     -.3126389    .0249212
engl_gpa |  -.3245378    .0497126     -6.528   0.000     -.4229026   -.2261729
   _cons |    .188824    .4686449      0.403   0.688     -.7384699    1.116118
------------------------------------------------------------------------------
(Outcome pvtq1_4==4 is the comparison group)
```

Rounding to three digits, the models we estimated are:

$$\log\left(\frac{\Pr(PVTQ1\_4=1)}{\Pr(PVTQ1\_4=4)}\right)=0.075+0.088*AGE\_KID-0.743*BIOMAPA-0.512*ENGL\_GPA$$

$$\log\left(\frac{\Pr(PVTQ1\_4=2)}{\Pr(PVTQ1\_4=4)}\right)=1.89-0.003*AGE\_KID-0.546*BIOMAPA-0.486*ENGL\_GPA$$

$$\log\left(\frac{\Pr(PVTQ1\_4=3)}{\Pr(PVTQ1\_4=4)}\right)=0.189+0.067*AGE\_KID-0.144*BIOMAPA-0.325*ENGL\_GPA$$

Understanding the Predictive Model

The following table shows the predictive models for each level of the BIOMAPA variable.

	Ln(Relative Risk) for model	
Model	Biomapa=0 (No)	Biomapa=1 (Yes)
1 to 4: <25th to ≥75th	$\beta_{0,1} + \beta_{1,1}$ AGE_KID + $\beta_{3,1}$ENGL_GPA	$\beta_{0,1} + \beta_{1,1} *$AGE_KID +$\beta_{2,1}$ + $\beta_{3,1}$ENGL_GPA
2 to 4: 25th - < 50th to ≥75th	$\beta_{0,2} + \beta_{1,2}$ AGE_KID + $\beta_{3,2}$ENGL_GPA	$\beta_{0,2} + \beta_{1,2} *$AGE_KID +$\beta_{2,2}$ + $\beta_{3,2}$ENGL_GPA
3 to 4: 50th - < 75th to ≥75th	$\beta_{1,3} + \beta_{1,3}$ AGE_KID + $\beta_{3,3}$ENGL_GPA	$\beta_{1,3} + \beta_{1,3} *$AGE_KID +$\beta_{2,3}$ + $\beta_{3,3}$ENGL_GPA

To determine the relative risk ratio of scoring under the 25th percentile to 75th percentile & over for adolescents living with both biological parents, you would compute:

$$\frac{e^{b_{0,1}+b_{1,1}AGE\_KID+b_{2,1}+b_{3,1}ENGL\_GPA}}{e^{b_{0,1}+b_{1,1}AGE\_KID+b_{3,1}ENGL\_GPA}} = e^{b_{2,1}} = e^{-0.74} = 0.48$$

We have considered the risk groups in the numerator and denominator to have the same age and grade in English. Thus, the risk is only 0.48 times as high of scoring under the 25th percentile versus over the 75th percentile if a adolescent lives with both biological parents. Computing the same risk ratio for the 50th to <75th score category compared to the 75th & over category, we find adolescents not living with both biological parents have a risk that is $e^{-0.14}$ = 0.87 times as those that do live with both biological parents. Note that these match the values in the Table of odds ratios printed by SUDAAN.

SURVIVAL (TIME-TO-EVENT) ANALYSIS

Survival analysis is used to predict the occurrence and timing of events. An event marks a qualitative change in status of the person (or entity) you are observing and the time that the event occurred. Some examples in the Add Health data are time to first intercourse and duration of a relationship.

The example used in this section investigates time until dropping out of school. In the Wave II In-home questionnaire, participants who were no longer in school were asked to identify the date they last attended school and the reason they were no longer attending. The goal here will be to use a participants' status at Wave I to predict if they will drop out of school by the time of the Wave II interview. The variables in the following table were constructed from Wave II data and merged with the Wave I data:

Use	Variable Name	Meaning
Outcome Variables	OVER_14	Adolescent is 15 or older 0=14 or younger 1=15 or older
	DROPOUT	Adolescent dropped out of school 0=(includes adolescents still enrolled, graduates & adolescents who left for other reasons) 1=Yes, reason for leaving was drop out
	TIME_MON	Continuous Time in months Adolescent attended school since Wave I interview
Design Variable	GSWGT2	Grand Sample Weight for wave II

Here we have a choice:

- include all of Wave I Respondents and use the Wave I design information
- include only the Wave II respondents and use the Wave II design information.

The analysis was ran both ways and identical results (within ± 0.01) were computed. Only the Wave II results are presented here.

Research Question

Do any of the following factors: race, sex of respondent, living with both biological parents, smoking, being held back a grade, trouble getting along with teachers, raise (or lower) an adolescent's hazard of dropping out of school? We will limit the analysis to adolescents over the age of 14.

Predictive Model

$$\log\frac{h(t)}{h_o(t)} = b_1 BLACK\_ + b_2 WHITE\_ + b_3 HISPANIC + b_4 BIOMAPA + b_5 MALE + b_6 SMKCIG + b_7 HELDBACK$$

where

h(t) = Hazard rate at time t, for the values of the covariates

$h_o(t)$ = Baseline hazard rate at time t, for values of all covariates equal to 0

β_1 = Hazard ratio for black race compared to not black

β_2 = Hazard ratio for white race compared to not white

β_3 = Hazard ratio for hispanic race compared to not hispanic

β_4 = Hazard ratio for living with both biological parents versus not living with biological parents

β_5 = Hazard ratio for males versus females

β_6 = Hazard ratio for smoking cigarettes regularly versus not smoking regularly

β_7 = Hazard ratio for being held back a grade.

Hazard is the probability that an adolescent will drop out of school in a month (the unit time interval) given that the adolescent has remained in school up to a given point time. Hence hazard is the rate of change of probability and can have values from 0 to infinity.

Using Stata: stcox

Stata does not have a special command for survival analysis with survey data, so we will use stset with the *pweight* option and stcox with *robust cluster()* option.

Stata Code

The *stset* command defines the failure time variable (*time_mon),* the grand sample weight variable (*gswgt2),* and the failure event (*dropout)* for use in subsequent survival commands. The *stcox* command specifies the covariates in the model. The *if* command subsets the data, limiting analysis to adolescents that are over 14 at time of the Wave I interview. The *robust cluster()* option specifies the primary sampling unit (*psuscid*).

```
stset time_mon [pweight=gswgt2], failure(dropout=1)
stcox age_kid black_ white_ hispanic biomapa male
       smkcig backgrad trbteach if over_14==1 , robust
       cluster(psuscid)
```

Stata Output

The *stset* command sets up key variables for survival analysis failure time (*time_mon*), grand sample weight (*gswgt2),* and failure event (*dropout*). Stata correctly reports 13,570 participants in Wave II, 63 participants with missing failure time data, and 219 participants who had dropped out of school before Wave I. Hence these adolescents will be omitted from any analyses. There are 275 adolescents who dropped out between Wave I and Wave II, while 13,288 adolescents were still in school or had left for other reasons.

```
failure event:  dropout == 1
obs. time interval:  (0, time_mon]
 exit on or before:  failure
            weight:  [pweight=gswgt2]
--------------------------------------------------------------------------------
    13570  total obs.
       63  event time missing (time_mon==.)                    PROBABLE ERROR
      219  obs. end on or before enter()
--------------------------------------------------------------------------------
    13288  obs. remaining, representing
      275  failures in single record/single failure data
 143622.8  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  16.26667
```

The *stcox* procedure produced the rest of the output. There were 7,702 adolescents over 14 with data included in this analysis. Below are the estimated hazard ratios from the fitted model. We see adolescents who are regular smokers (*smkcig*) or who have been held back a grade in school (*backgrad*) have risks of dropping out of school approximately three times greater than other adolescents. Age (*age_kid)* and having trouble with teachers (*trbteach*) also elevates the hazard of dropping out of school, while adolescents who live with both biological parents (*biomapa*) have only 0.47 times the risk. Sex of adolescent (*male)* and

selecting white (*white_*) race has no effect. Selecting black race (p=0.061) or hispanic (p=0.088) ethnic group are marginally significant.

```
        failure _d:  dropout == 1      analysis time _t:  time_mon
            weight:  [pweight=gswgt2]  Cox regression -- Breslow method for ties

No. of subjects =  9507222.812                     Number of obs   =      7702
No. of failures =  394209.4104
Time at risk    =  100210633.9                     Wald chi2(9)    =    291.76
Log likelihood  =   -2571.3805                     Prob > chi2     =    0.0000
                          (standard errors adjusted for clustering on psuscid)
------------------------------------------------------------------------------
      _t |               Robust
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 age_kid |   1.471376   .1107276      5.132   0.000         1.2696     1.70522
  black_ |   .5291449   .1795304     -1.876   0.061       .2721295    1.028901
  white_ |   1.038147   .1895911      0.205   0.838       .7257855     1.48494
hispanic |   1.413141   .2864434      1.706   0.088       .9498328    2.102441
 biomapa |   .4737559   .0820715     -4.312   0.000       .3373618    .6652935
    male |   .8495824   .1554353     -0.891   0.373       .5935728     1.21601
  smkcig |   3.239976   .6366865      5.982   0.000       2.204309    4.762238
backgrad |   2.919306   .5900302      5.301   0.000       1.964448    4.338291
trbteach |    1.42784   .1125872      4.517   0.000       1.223379    1.666472
```

APPENDIX A. STATA CODE TEMPLATES

Stata Code Template

The sampling characteristics are defined with the *svyset* statement:

```
svyset strata region
svyset pweight wt_var
svyset psu psuscid
```

Use the *strata* keyword to specify the stratification variable (*region)*, the *pweight* keyword to specify the weight variable (*wt_var)*, and the *psu* keyword to specify the primary sampling unit (*psuscid)*. Stata defaults to a "With Replacement" design type. Next, use one of the Stata commands for survey analysis. These commands begin with the letters svy. For example, to compute mean PVT scores you would use:

```
svymean ah_pvt
```

Subpopulation analysis is done by using either the *subpop* or *by* option. For example, to do the above analysis for boys you could create an indicator variable called *sex* with the value 1 for boys and 0 for girls and then use:

```
svymean ah_pvt, subpop(sex)
```

or use the by statement

```
svymean ah_pvt, by sex
```

Stata provides the *svytest* command for customized hypothesis testing and *svylc* for estimating linear combinations of parameter estimates after estimation with any of the survey commands.

Stata Robust Variance Estimation

Certain stata commands allow you to specify *pweights* (probability weights) and the *robust cluster* option. Below is an example using the logit command. Note for the commands that have the robust cluster option, sub-population analysis is done via the *if* statement.

```
logit outcome var1 var2 var3 if group==1
       [pweight=wt_var], robust cluster(psuscid)
```

Linear combinations of parameter estimates can be obtained with the *lincom* command; hypothesis tests can be performed with the *test* command.

REFERENCES

Brogan, D., Daniels, D., Rolka, D. Marsteller, F. Chattopadhay, M., "Software for Sample Survey Data: Misuse of Standard Packages"; invited chapter in Encyclopedia of Biostatistics, editors-in-chief Peter Armitage and Theodore Colton, John Wiley, New York, Volume 5, 1998, pages 4167-4174.

Chantala, K. and Tabor, J. "National Longitudinal Study of Adolescent Health Strategies to Perform a Design-Based Analysis Using the Add Health Data" University of North Carolina at Chapel Hill, 1999.

Cohen, S. B., "An Evaluation of Alternative PC-Based Software Packages Developed for the Analysis of Complex Survey Data," The American Statistician, August 1997, Vol. 51, No. 3, pages 285-292.

Levy, P. S., and Lemeshow, S., "Sampling of Populations Methods and Applications," John Wiley & Sons, 1999. 525 p.

SAS Institute Inc., "SAS/STAT Software: Changes & Enhancements through Release 6.12," Cary, NC, SAS Institute, 1997.

Shah, B. V., Barnwell, B. G., and Bieler, G. S., "SUDAAN User's Manual: Release 6.4," Research Triangle Institute, Research Triangle Park, NC, 1995.

Stata Corporation, "Stata Reference Manual," Release 6, College Station, TX, 1999.

Touraneau, R. and Hee-Choon, S. "National Longitudinal Study of Adolescent Health Grand Sample Weight," Carolina Population Center, University of North Carolina at Chapel Hill