CHAPTER 10 ASDA ANALYSIS EXAMPLES REPLICATION-SAS V9.2

GENERAL NOTES ABOUT ANALYSIS EXAMPLES REPLICATION

These examples are intended to provide guidance on how to use the commands/procedures for analysis of complex sample survey data and assume all data management and other preliminary work is done. The relevant syntax for the procedure of interest is shown first along with the associated output for that procedure(s). In some examples, there may be more than one block of syntax and in this case all syntax is first presented followed by the output produced.

In some software packages certain procedures or options are not available but we have made every attempt to demonstrate how to match the output produced by Stata 10+ in the textbook. Check the ASDA website for updates to the various software tools we cover.

GENERAL NOTES ON SURVIVAL ANALYSIS

This chapter demonstrates how to use PROC LIFETEST and PROC SURVEYLOGISTIC for descriptive survival curves and discrete time logistic regression with binary outcomes. The discrete time logistic regression example could be generalized to other types of logistic regression such as multinomial and ordinal regression as well. Some fine points of use of PROC LIFETEST include use of the FREQ statement with integer weights (created by multiplying the weight by a large constant such as 1000 in the data step), saving key statistics for graphing with PROC GPLOT, and use of a variable which represents age of event or age at censoring in the time statement.

For the discrete time logistic models the syntax is the same as for a non-survival type of model but the data set must be structured to include multiple records per individual to correctly account for the time concept within the models. For the comparison between logistic and Clog-log regression use of the link=cloglog option on the model statement is used. Because the link=cloglog default output does not include odds ratios, additional code to output the parameter estimates and perform exponentiation is included here.

Some options to note: use of the class statement requires the / param=ref specification if you want to use a reference group parameterization instead of the default effects coding approach, use of the (ref=first) allows specification of the omitted category for the class variables, and use of the (event='1') syntax declares the probability modeled for the outcome variable. There are examples of these options in this chapter. PROC SURVEYLOGISTIC also allows the use of the test statement and use of the crossing operator for interaction variables in the model statement.

SAS v9.2 does not offer a complex sample survey procedure for the Cox PH models and thus these examples are omitted here.

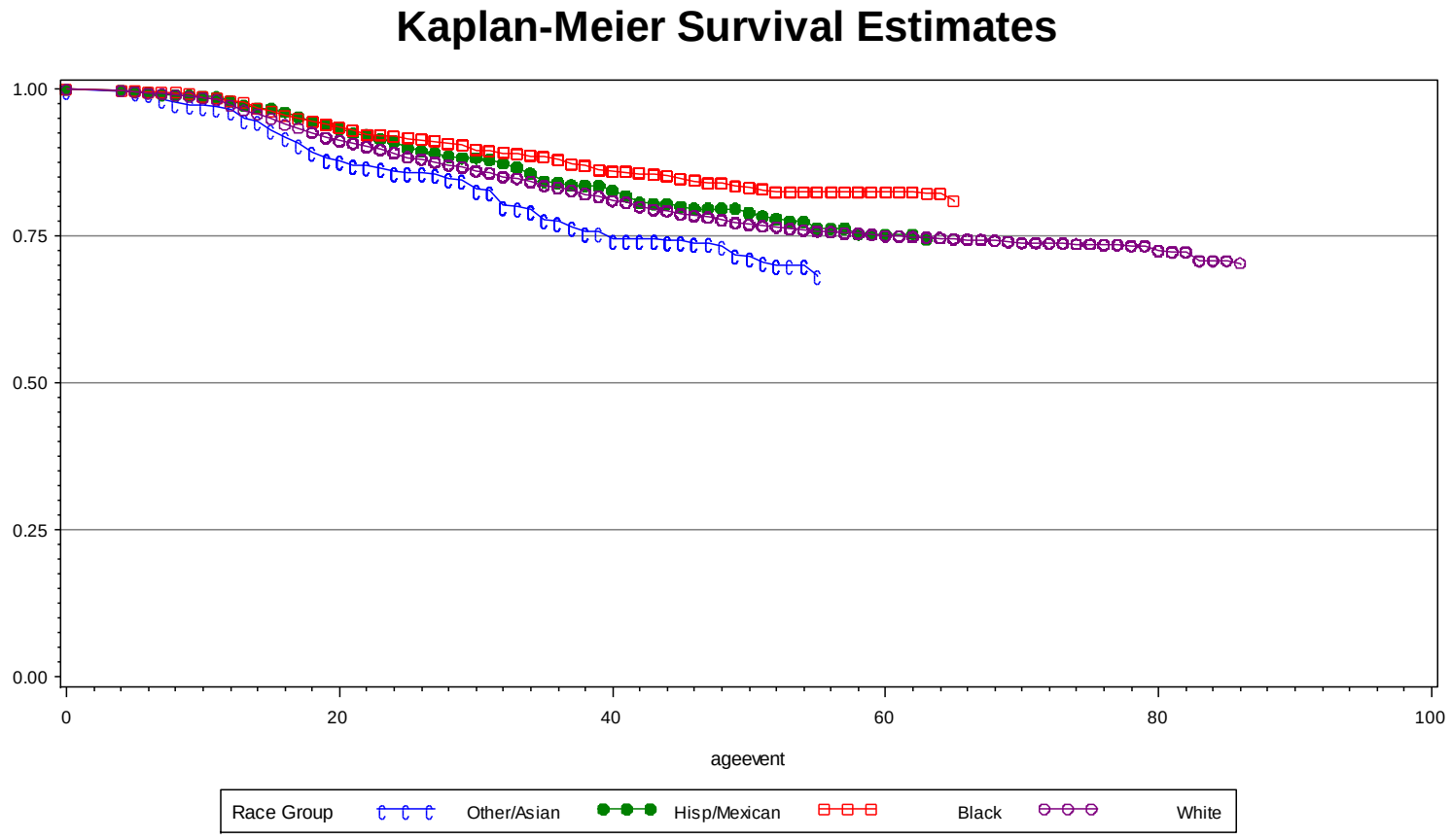
```
ANALYSIS EXAMPLE OF KAPLAN-MEIER CURVE NCS-R DATA

* note that you must use a freq statement rather than a weight statement in PROC LIFETEST: this means use of a
        non-integer freq weight rather than the usual weight statement that allows an integer weight ;
*note that PROC LIFETEST does include built in graphics but use of PROC GPLOT provides more flexibility in producing custom graphics;

proc lifetest data=c10  outsurv=survival ;
title "Kaplan-Meier Survival Estimates by Race " ;
time ageevent*mde(0) ;
strata racecat ;
freq ncsrwtlg100 ;
run ;

proc format ;
value rf 1='Other/Asian' 2='Hisp/Mexican' 3='Black' 4='White' ;
run ;

*note that you could use the plots option in proc lifetest but use of proc gplot allows
Control over the lines and axes for the final graph ;
symbol1 v=cross c=blue i=j ;
symbol2 v=dot c=green i=j ;
symbol3 v=square c=red i=j;
symbol4 v=circle c=purple i=j  ;
axis1  order=0 to 1.00 by .25 label=none ;
axis2 order = 0 to 100 by 20 ;
legend1 label=('Race Group') frame ;
proc gplot data=survival ;
plot survival*ageevent =stratum /
        vaxis=axis1 haxis=axis2 legend=legend1 vref=(.25 .5 .75) ;
format stratum rf. ;
title "Kaplan-Meier Survival Estimates" ;
run ;
```



NOTE: SAS DOES NOT OFFER THE SURVEY CORRECTED COX PROPORTIONAL HAZARDS MODEL (NOT INCLUDED HERE)

ANALYSIS EXAMPLE: DISCRETE TIME LOGISTIC REGRESSION WITH A LOGIT LINK NCS-R DATA
NOTE: INT REPSENTS PERSON YEARS AND IS DESCRIBED AS PYR IN THE ASDA TEXT.

```
*Note that the data set has previously been structured to create a multiple record per person file for these analyses!
        note that the where statement includes records up to and including the year of the event of interest if occurred or year
        censored if no event of interest ;

title "Discrete Time Logistic Regression" ;
proc surveylogistic data=ncsrpersonyear order=data ;
strata sestrat ;
cluster seclustr ;
weight ncsrwtsh ;
class ed4cat (ref=first) racecat (ref=first) mar3cat (ref=first) sex (ref=last) / param=ref ;
model mdetv (event='1') =ed4cat racecat mar3cat sex int intwage ;
where int <= ageonsetmde ;
format ed4cat edf. racecat rf. mar3cat marf. sex sf. ;
run ;
```

Discrete Time Logistic Regression
The SURVEYLOGISTIC Procedure

Model Information

Data Set	WORK.NCSRPERSONYEAR	
Response Variable	mdetv	
Number of Response Levels	2	
Stratum Variable	SESTRAT	SAMPLING ERROR STRATUM
Number of Strata	42	
Cluster Variable	SECLUSTR	SAMPLING ERROR CLUSTER
Number of Clusters	84	
Weight Variable	NCSRWTSH	NCSR sample part 1 weight
Model	Binary Logit	
Optimization Technique	Fisher's Scoring	
Variance Adjustment	Degrees of Freedom (DF)	

Variance Estimation

Method	Taylor Series
Variance Adjustment	Degrees of Freedom (DF)
Number of Observations Read	385696
Number of Observations Used	385696
Sum of Weights Read	386866
Sum of Weights Used	386866

Response Profile

Ordered Value	mdetv	Total Frequency	Total Weight
1	0	383867	385086.58
2	1	1829	1779.46

Probability modeled is mdetv=1.

Class Level Information

Class	Value	Design Variables		
ED4CAT	1=0-11 Yrs	0	0	0
	2=12 Yrs	1	0	0
	3=11-13 Yrs	0	1	0
	4=16+ Yrs	0	0	1
racecat	1=Other/Asian	0	0	0
	2=Hisp/Mexican	1	0	0
	3=Black	0	1	0
	4=White	0	0	1
MAR3CAT	1=Married	0	0	
	2=Previously Married	1	0	
	3=Never Married	0	1	
SEX	1=Male	1		
	2=Female	0		

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	22706.046	21681.912
SC	22716.909	21812.265
-2 Log L	22704.046	21657.912

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1046.1350	11	<.0001
Score	927.1033	11	<.0001
Wald	774.3137	11	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ED4CAT	3	5.8718	0.1180
racecat	3	37.6833	<.0001
MAR3CAT	2	70.3758	<.0001
SEX	1	51.0086	<.0001
Int/pyr	1	250.1251	<.0001
intwage	1	567.5308	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.4355	0.1620	449.7932	<.0001
ED4CAT	2=12 Yrs	1	-0.0201	0.0661	0.0928	0.7607
ED4CAT	3=11-13 Yrs	1	0.0929	0.0574	2.6163	0.1058
ED4CAT	4=16+ Yrs	1	-0.0195	0.0633	0.0943	0.7587
racecat	2=Hisp/Mexican	1	-0.2484	0.1348	3.3977	0.0653
racecat	3=Black	1	-0.4570	0.1499	9.2945	0.0023
racecat	4=White	1	0.0740	0.1182	0.3916	0.5314
MAR3CAT	2=Previously Married	1	0.4942	0.0610	65.6261	<.0001
MAR3CAT	3=Never Married	1	-0.0353	0.0880	0.1614	0.6879
SEX	1=Male	1	-0.4449	0.0623	51.0086	<.0001
Int/pyr		1	0.0328	0.00207	250.1251	<.0001
intwage		1	-0.0583	0.00245	567.5308	<.0001

Odds Ratio Estimates

Effect			Point Estimate	95% Wald Confidence Limits	
ED4CAT	2=12 Yrs	vs 1=0-11 Yrs	0.980	0.861	1.116
ED4CAT	3=11-13 Yrs	vs 1=0-11 Yrs	1.097	0.981	1.228
ED4CAT	4=16+ Yrs	vs 1=0-11 Yrs	0.981	0.866	1.110
racecat	2=Hisp/Mexican	vs 1=Other/Asian	0.780	0.599	1.016
racecat	3=Black	vs 1=Other/Asian	0.633	0.472	0.849
racecat	4=White	vs 1=Other/Asian	1.077	0.854	1.358
MAR3CAT	2=Previously Married	vs 1=Married	1.639	1.455	1.847
MAR3CAT	3=Never Married	vs 1=Married	0.965	0.812	1.147
SEX	1=Male vs 2=Female		0.641	0.567	0.724
int			1.033	1.029	1.038
intwage			0.943	0.939	0.948

Association of Predicted Probabilities and Observed Responses

Percent Concordant	63.6	Somers' D	0.406
Percent Discordant	23.1	Gamma	0.468
Percent Tied	13.3	Tau-a	0.004
Pairs	702092743	c	0.703

```
DISCRETE TIME LOGISTIC REGRESSION WITH A CLOGLOG LINK FOR COMPARISON TO LOGIT LINK
NOTE: NEED TO SAVE PARAMETER ESTIMATES IN OUTPUT DATA SET AND THEN CREATE ODDS RATIOS USING A VARIABLE CREATION PROCESS

*USE OF ODS OUTPUT OPTION TO SAVE PARAMETER ESTIMATES ;

ods output parameterestimates=estimates ;
title "CLogLog Regression to Compare with Logistic Regression" ;
proc surveylogistic data=ncsrpersonyear ;
strata sestrat ;
cluster seclustr ;
weight ncsrwtsh ;
class ed4cat (ref=first) racecat (ref=first) mar3cat (ref=first) sex (ref=last) / param=ref ;
model mdetv (event='1') =ed4cat racecat mar3cat sex int intwage / link=cloglog ;
where int <= ageonsetmde ;
format ed4cat edf. racecat rf. mar3cat marf. sex sf. ;
run ;
ods output close ;

CLogLog Regression to Compare with Logistic Regression
The SURVEYLOGISTIC Procedure
                              Model Information
Data Set                        WORK.NCSRPERSONYEAR
Response Variable               mdetv
Number of Response Levels       2
Stratum Variable                SESTRAT                     SAMPLING ERROR STRATUM
Number of Strata                42
Cluster Variable                SECLUSTR                    SAMPLING ERROR CLUSTER
Number of Clusters              84
Weight Variable                 NCSRWTSH                    NCSR sample part 1 weight
Model                           Binary Cloglog
Optimization Technique          Fisher's Scoring
Variance Adjustment             Degrees of Freedom (DF)


             Variance Estimation

Method                             Taylor Series
Variance Adjustment    Degrees of Freedom (DF)


Number of Observations Read      385696
Number of Observations Used      385696
Sum of Weights Read              386866
Sum of Weights Used              386866

                  Response Profile

 Ordered                       Total            Total
   Value        mdetv     Frequency           Weight

       1            0        383867        385086.58
       2            1          1829          1779.46

Probability modeled is mdetv=1.

               Class Level Information

Class       Value                    Design Variables

ED4CAT      1=0-11 Yrs                0      0      0
            2=12 Yrs                  1      0      0
            3=11-13 Yrs               0      1      0
            4=16+ Yrs                 0      0      1

racecat     1=Other/Asian             0      0      0
            2=Hisp/Mexican            1      0      0
            3=Black                   0      1      0
            4=White                   0      0      1

MAR3CAT     1=Married                 0      0
            2=Previously Married      1      0
            3=Never Married           0      1

SEX         1=Male                    1
            2=Female                  0
```

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

CLogLog Regression to Compare with Logistic Regression
The SURVEYLOGISTIC Procedure

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	22706.046	21681.795
SC	22716.909	21812.148
-2 Log L	22704.046	21657.795

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1046.2516	11	<.0001
Score	927.1033	11	<.0001
Wald	774.9840	11	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ED4CAT	3	5.8284	0.1203
racecat	3	37.6263	<.0001
MAR3CAT	2	70.5804	<.0001
SEX	1	50.9709	<.0001
Int/pyr	1	250.2851	<.0001
intwage	1	568.3351	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.4444	0.1614	455.5649	<.0001
ED4CAT	2=12 Yrs	1	-0.0197	0.0659	0.0899	0.7643
ED4CAT	3=11-13 Yrs	1	0.0924	0.0572	2.6070	0.1064
ED4CAT	4=16+ Yrs	1	-0.0192	0.0631	0.0926	0.7608
racecat	2=Hisp/Mexican	1	-0.2474	0.1344	3.3907	0.0656
racecat	3=Black	1	-0.4551	0.1494	9.2731	0.0023
racecat	4=White	1	0.0737	0.1179	0.3912	0.5316
MAR3CAT	2=Previously Married	1	0.4928	0.0608	65.7626	<.0001
MAR3CAT	3=Never Married	1	-0.0355	0.0875	0.1642	0.6853
SEX	1=Male	1	-0.4432	0.0621	50.9709	<.0001
Int/pyr		1	0.0327	0.00207	250.2851	<.0001
intwage		1	-0.0582	0.00244	568.3351	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	63.6	Somers' D	0.405
Percent Discordant	23.1	Gamma	0.467
Percent Tied	13.3	Tau-a	0.004
Pairs	702092743	c	0.703

```
data estimates_exp ;
        set estimates ;
oddsratios=exp(estimate) ;
proc print data=estimates_exp ;
run ;
```

CLogLog Regression to Compare with Logistic Regression

Obs	Variable	ClassVal0	DF	Estimate	StdErr	WaldChiSq	Prob ChiSq	oddsratios
1	Intercept		1	-3.4444	0.1614	455.5649	<.0001	0.03192
2	ED4CAT	2=12 Yrs	1	-0.0197	0.0659	0.0899	0.7643	0.98045
3	ED4CAT	3=11-13 Yrs	1	0.0924	0.0572	2.6070	0.1064	1.09676
4	ED4CAT	4=16+ Yrs	1	-0.0192	0.0631	0.0926	0.7608	0.98098
5	racecat	2=Hisp/Mexican	1	-0.2474	0.1344	3.3907	0.0656	0.78081
6	racecat	3=Black	1	-0.4551	0.1494	9.2731	0.0023	0.63440
7	racecat	4=White	1	0.0737	0.1179	0.3912	0.5316	1.07652
8	MAR3CAT	2=Previously Married	1	0.4928	0.0608	65.7626	<.0001	1.63691
9	MAR3CAT	3=Never Married	1	-0.0355	0.0875	0.1642	0.6853	0.96515
10	SEX	1=Male	1	-0.4432	0.0621	50.9709	<.0001	0.64197
11	int/pyr		1	0.0327	0.00207	250.2851	<.0001	1.03327
12	intwage		1	-0.0582	0.00244	568.3351	<.0001	0.94348