NCS-Replication Analysis Training Program
July 17-19, 2006

- Three day session covering analysis process including obtaining data and documentation, selection of variables needed for analysis, preparation for analysis and performing common analyses with NCS-R data

- All analyses will be demonstrated using SAS v9.1 and hands-on computing practice using SAS is built into training session

- Focus will be big-picture of typical analyses and how to perform using SAS

- We will not focus on statistics or how to code in SAS, assume prior knowledge of stats and software of choice, those that do not use SAS will be provided with programs to run

- Discussion will attempt to cover common analysis tasks so that participants can generalize to their own work

NCS-Replication Analysis Training Program
July 17-19, 2006

Monday July 17, 2006

Morning Session-9am-Noon – Nancy Sampson

- Presentation and Discussion of Diagnostic Variables

Afternoon Session - 1-5 Pm – Pat Berglund

- Overview of NCS-R websites, NCS-R datasets, documentation products
- Overview of ICPSR/SAMDHA Website including SDA online analysis system - JoAnne McFarland O'Rourke
- NCS-R Instrument - Part 1 and Part 2, Sub-Samples
- Discussion of Sample Design, Weights and Analysis of Complex Data
- Introduction to SAS v9.1

NCS-Replication Analysis Training Program
July 17-19, 2006

Tuesday July 18 – Analysis Planning, NCS-R Analysis Perspectives, Descriptive Statistics, Regression

Morning Session – 9-11am – Pat Berglund and Mike Gruber

- Planning an NCS-R analysis – instrument sections, variable creation, weights, analysis techniques
- Exploration of NCS-R datasets via printouts, contents, descriptive tables
- Means, proportions, and univariate analyses, subgroup analysis in SAS, corrected standard errors
- 11-12:30 - Discussion of NCS-R Analysis Topics and Historical Perspectives – Dr. Kessler

Tuesday July 18 – 1:30-5pm - Pat Berglund and Mike Gruber

- Production of typical descriptive analyses including demographic prevalences and diagnostic prevalences by gender using SAS v9.1
- Analysis of NCS-R data using logistic regression and other common modeling techniques, hands-on computer work doing regression
- Individual meetings with Dr. Kessler regarding analysis interests/projects

- Optional Group Dinner at Szechwan West Restaurant @ 7:30pm – details to follow

NCS-Replication Analysis Training Program
July 17-19, 2006

Wednesday July 19, 2006

Morning Session- 9am-noon- Pat Berglund and Mike Gruber

Survival Curves and Data Preparation for Survival Analysis

- Preparation of data for survival analysis
- Discussion and demonstration of SAS proc lifetest to produce survival curves
- Creating person-year files, time-varying covariates and time dependent outcomes for survival analysis via discrete-time logistic regression

Afternoon Session – 1pm-5pm – Pat Berglund and Mike Gruber

Discrete-Time Logistic Regression

- Demonstration of discrete time logistic regression in SAS, replication of survival model results including hands-on analysis using SAS

- General Question and Answer Period with NCS-R Analysts

Overview of NCSR Web-Based Tools

- Use the NCS-R website as a good starting point for NCS-R information
- www.hcp.med.harvard.edu/ncs
- This site includes links to instruments, publications, FAQ, and other key tools for the analyst
- Other useful sites might be the World Mental Health Initiative website, and software sites such as sas.com, Stata homepage and SPSS site

National Comorbidity Survey (NCS) and National Comorbidity Survey Replication (NCS-R)

Click here to access the public release of the NCS-R dataset and find information about our training workshops.

The baseline NCS, fielded from the fall of 1990 to the spring of 1992, was the first nationally representative mental health survey in the U.S. to use a fully structured research diagnostic interview to assess the prevalences and correlates of DSM-III-R disorders. The baseline NCS respondents were reinterviewed in 2001-02 (NCS-2) to study patterns and predictors of the course of mental and substance use disorders and to evaluate the effects of primary mental disorders in predicting the onset and course of secondary substance disorders. In conjunction with this, an NCS Replication survey (NCS-R) was carried out in a new national sample of 10,000 respondents. The goals of the NCS-R are to study trends in a wide range of variables assessed in the baseline NCS and to obtain more information about a number of topics either not covered in the baseline NCS or covered in less depth than we currently desire. A survey of 10,000 adolescents (NCS-A) was carried out in parallel with the NCS-R and NCS-2 surveys. The goal of NCS-A is to produce nationally representative data on the prevalences and correlates of mental disorders among youth. The NCS-R and NCS-A, finally, are being replicated in a number of countries around the world. Centralized cross-national analysis of these surveys is being carried out by the NCS data analysis team under the auspices of the World Health Organization (WHO) World Mental Health Survey Initiative.

In order to provide an easily accessible database which can be updated and checked on a regular basis, we have created a public use file system containing all the documents from the NCS Program. This file system can be accessed through the Internet and either downloaded onto a disk or printed. We will update the system on a regular basis to add newly completed paper abstracts and other documents. In addition, the NCS data can be accessed through ICPSR (Inter-university Consortium for Political and Social Research). Any updates to the data to correct coding or

Overview of Public Release Dataset

- All raw and selected diagnostic, demographic, sample design and weight variables are included in one file, n=9282 for the entire part 1 sample with a sub-sample of respondents that completed part 2 of the instrument, n=5692 part 2 respondents
  - Raw variables-includes all variables that could be released while keeping disclosure issues in mind
  - Diagnostic variables-includes selected diagnostic variables such as ICD and DSM disorders along with age of onset, age of recency, lifetime, 12 month disorders, and 30 day disorders
  - Demographic variables-includes selected demographic and design variables
- The dataset can be downloaded in various formats such as SAS transport, ASCII with SAS, SPSS, or Stata setup statements
- Associated documentation tools include: online and Adobe/pdf format codebook, and Adobe/pdf format of the instrument
- Other tools are related literature links and background information on the study and analysis tips including sample programs (in the pdf version of the codebook)

ICPSR Site Tour

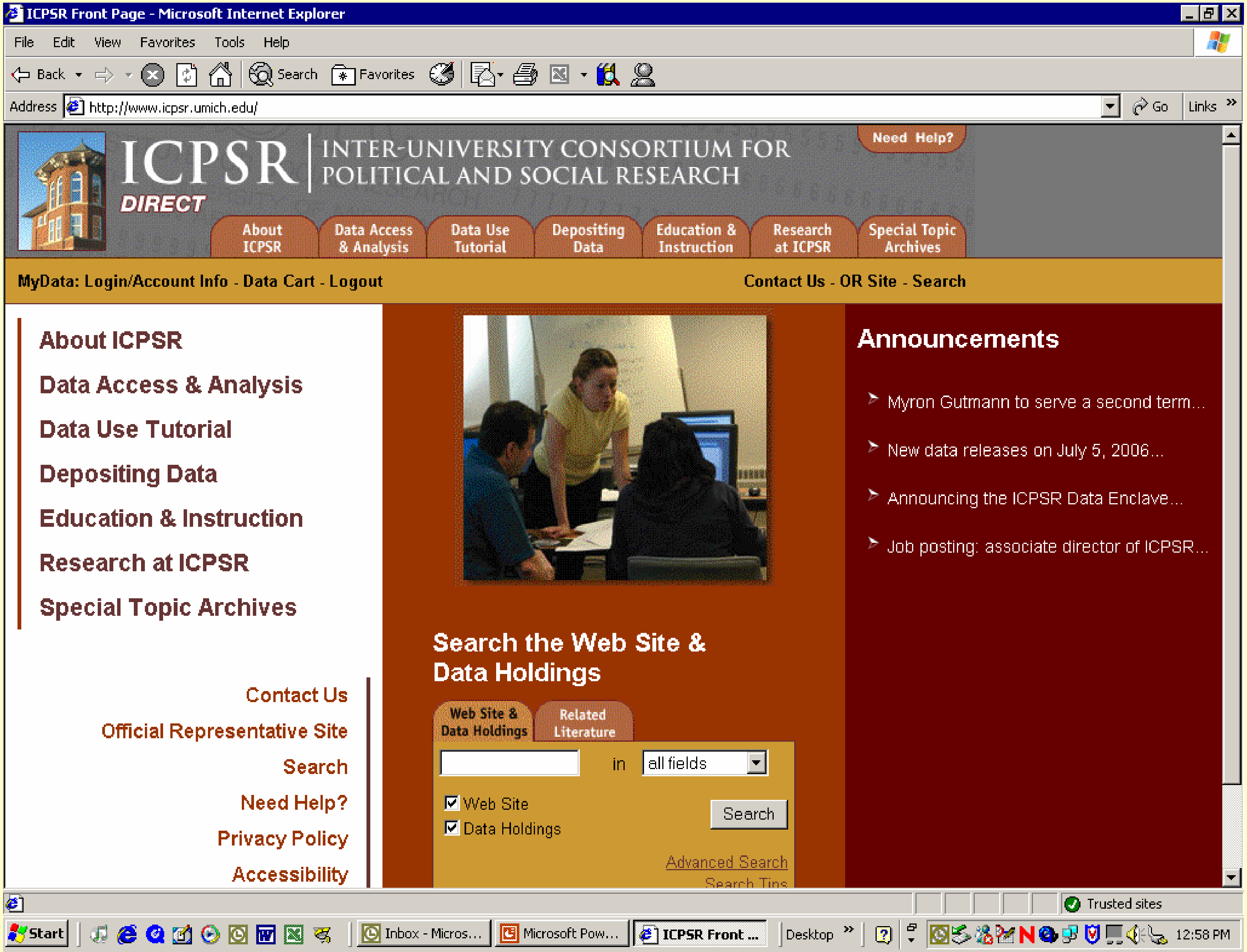
- We are fortunate to have JoAnne McFarland-O'Rourke, Principal Investigator of SAMSHA, willing to guide us through the ICPSR/NCS-R site

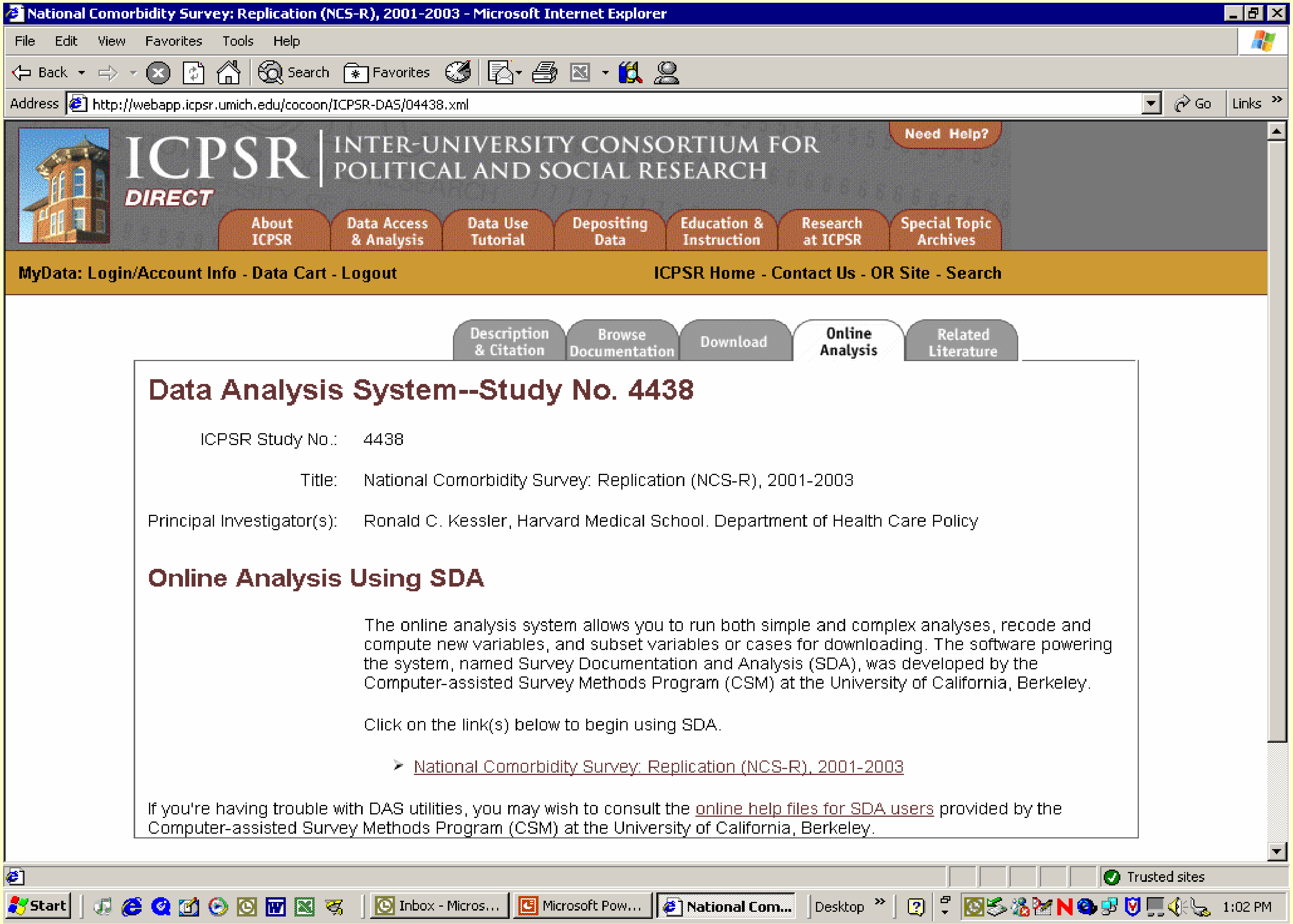
- ICPSR and SAMSHA have been instrumental in providing documentation and archive tools for the NCS, and the NCS-R

- ICPSR also provides an online analysis tool for both the NCS and the NCS-R datasets

ICPSR - NCS Surveys Website

- The ICPSR site is another key starting point for downloading and analyzing all NCS datasets (NCS, NCS-R)
- The site includes the public release data in various formats, codebook in both pdf and online formats, the instrument in pdf format and links to related literature
- The site also offers an online analysis system for the NCSR dataset and offers an easy and quick way to obtain results from the NCS-R public release file


ICPSR Front Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.icpsr.umich.edu/
Go Links
ICPSR
DIRECT
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
Need Help?
About ICPSR
Data Access & Analysis
Data Use Tutorial
Depositing Data
Education & Instruction
Research at ICPSR
Special Topic Archives
MyData: Login/Account Info - Data Cart - Logout
Contact Us - OR Site - Search
About ICPSR
Data Access & Analysis
Data Use Tutorial
Depositing Data
Education & Instruction
Research at ICPSR
Special Topic Archives
Contact Us
Official Representative Site
Search
Need Help?
Privacy Policy
Accessibility
Search the Web Site & Data Holdings
Web Site & Data Holdings
Related Literature
in
all fields
Web Site
Data Holdings
Search
Advanced Search
Search Tips
Announcements
Myron Gutmann to serve a second term...
New data releases on July 5, 2006...
Announcing the ICPSR Data Enclave...
Job posting: associate director of ICPSR...
Trusted sites
Start
Inbox - Micros...
Microsoft Pow...
ICPSR Front ...
Desktop
12:58 PM


National Comorbidity Survey: Replication (NCS-R), 2001-2003 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://webapp.icpsr.umich.edu/cocoon/ICPSR-DAS/04438.xml
Go Links
ICPSR DIRECT
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH
Need Help?
About ICPSR
Data Access & Analysis
Data Use Tutorial
Depositing Data
Education & Instruction
Research at ICPSR
Special Topic Archives
MyData: Login/Account Info - Data Cart - Logout
ICPSR Home - Contact Us - OR Site - Search
Description & Citation
Browse Documentation
Download
Online Analysis
Related Literature
Data Analysis System--Study No. 4438
ICPSR Study No.: 4438
Title: National Comorbidity Survey: Replication (NCS-R), 2001-2003
Principal Investigator(s): Ronald C. Kessler, Harvard Medical School. Department of Health Care Policy
Online Analysis Using SDA
The online analysis system allows you to run both simple and complex analyses, recode and compute new variables, and subset variables or cases for downloading. The software powering the system, named Survey Documentation and Analysis (SDA), was developed by the Computer-assisted Survey Methods Program (CSM) at the University of California, Berkeley.
Click on the link(s) below to begin using SDA.
National Comorbidity Survey: Replication (NCS-R), 2001-2003
If you're having trouble with DAS utilities, you may wish to consult the online help files for SDA users provided by the Computer-assisted Survey Methods Program (CSM) at the University of California, Berkeley.
Trusted sites
Start
Inbox - Micros...
Microsoft Pow...
National Com...
Desktop
1:02 PM

NCS-R Instrument - Sections and Flow

- The NCS-R contains 46 sections, most have been released for the public version of the dataset, some such as dementia were omitted due to confidentiality issues

- The instrument is divided into 2 parts
  - Part 1 includes sections 1-14 with an additional demographic section for those that do not go on to complete Part 2
  - Part 2 includes detailed questions about additional disorders such as gambling disorder, childhood disorders such as conduct disorder and ADD, social networks, family history/risk factors and other detailed sections such as finances

- At the end of the Pharmacoepidemiology section, a series of questions directing flow into Part 2 of the survey are included, other key flow questions are included in the Screener section and these are referenced in the Pharmacoepi section as well

Section Flow and Part 1 and Part 2 of NCS-R

Section	
1. Household Listing	
2. Screening (SC)	
3. Depression (D)	
4. Mania (M)	
5. Irritable Depression (IR)	
6. Panic Disorder (PD)	
7. Specific Phobia (SP)	
8. Social Phobia (SO)	
9. Agoraphobia (AG)	
10. Generalized Anxiety Disorder (G)	
11. Intermittent Explosive Disorder (IED)	
12. Suicidality (SD)	
13. Services (SR)	
14. Pharmacoepidemiology (PH)	Long (100%) + Int (100%) + Short(100%)
15. Demographics (DM)	Short (100%)
16. Personality (PEA)	Long (100%) + Int (100%)
17. Substance Use (SU)	Long (100%)
18. Post-Traumatic Stress Disorder (PT)	Long (100%)
19. Chronic Conditions (CC)	Long (100%)
20. Neurasthenia (N)	Long (100%)
21. 30-Day Functioning (WHO-DAS)	Long (100%)
22. 30-Day Symptoms (NSD)	Long (100%) + Int (100%)
23. Tobacco (TB)	Long (100%)
24. Eating Disorders (EA)	Long (50%)
25. Premenstrual Syndrome (PR)	Long (100% females)
26. Obsessive-Compulsive Disorder (O)	Long (30%)
27. Psychosis (PS)	Long (30%)
28. Gambling (GM)	Long (50%)
29. Worries and Unhappiness (WU)	Long (30%) + Int (30%) + Short (30%)
30. Employment (EM)	Long (100%)+Int (100%)
31. Finances (FN)	Long (100%)+Int(100%)
32. Marriage (MR)	Long (100%)+Int(100%)
33. Children (CN)	Long (100%)+Int(100%)
34. Social Networks (SN)	Long (100%)+Int(100%)
35. Adult Demographics (DA)	Long (100%)+Int(100%)
36. Childhood Demographics (DE)	Long (100%)+Int(100%)
37. Childhood (CH)	Long (100%)+Int(100%)
38. Attention-Deficit/Hyperactivity (AD)	Long (100% of 44 yold and younger and stem)
39. Oppositional-Defiant Disorder (OD)	Long (100% of 44 yold and younger and stem)
40. Conduct Disorder (CD)	Long (100% of 44 yold and younger)
41. Separation Anxiety Disorder (SA)	Long (100% of stem)
42. Family Burden (FB)	Long (30%) + Int (30%) + Short (30%)
43. Perceptions of the Past (PP)	Long (25%) + Int (25%) + Short (25%)
44. Respondent Contacts	Long (100%) + Int (100) + Short (100%)
45. Interviewer's Observation (IO)	Long (100%) + Int (100) + Short (100%)
46. Dementia – PAPER ONLY	

Flow into Part 2 of Interview

- Questions ph100, ph101 and subsequent questions channel respondents to various sections of the questionnaire
- These questions are related to the screening questions of the Screener section (see instrument instructions)
- Overall strategy of using the Screener section is detailed in the paper "The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI)", RONALD C. KESSLER, T. BEDIRHAN ÜSTÜN

Rationale for 2 Parts to NCS-R Instrument

- Interview Length and Analysis content (see Kessler Design and Field paper)
- "The interview schedule was divided into two parts. Part I, administered to all respondents, included all core WMH-CIDI disorders. The administration time of Part I averaged 33.8 minutes and had an inter-quartile range between 22.6 and 39.8 minutes (Table 1).
- Part II included assessments of risk factors, consequences, services, and other correlates of the core disorders. Part II also included assessments of additional disorders that were either of secondary importance or that were very timeconsuming to assess.
- Part II was administered only to 5,692 of the 9,282 NCS-R respondents, over-sampling those with clinically significant psychopathology. All respondents who did not receive Part II were administered a brief demographic battery and were then either terminated or sampled in their appropriate proportions into sub-sampled interview sections that are described below."

Additional Sub-samples of the NCS-R

- Other subgroups are the "intermediate group" or couples sample:

***PH114**. INTERVIEWER CHECKPOINT: (SEE **RESPONDENT'S ID NUMBER**)

R IS PART OF COUPLES SAMPLE	1 **THESE ARE "INTERMEDIATE GROUP" – GO TO *PEA1, PAGE X**
ALL OTHERS	2 **THESE ARE "SHORT GROUP" – GO TO *DM1, NEXT SECTION**

- Additional sub-samples:
  - Disorders asked of those <=44 years old (Impulse disorders such as Conduct Disorder, ADD, etc)
  - Disorders asked only of certain subsets of the part 2 (Gambling and Eating Disorders 50% of the part 2)

NCS-R Sample Design

- General overview of sample design for NCSR
- What are elements of design?
- How is sample executed?
- Implications for analysis

NCSR Sample Design

- Nationally representative multi-stage clustered area probability sample of households. Interviewed people in the age range 18 years and older, rather than in the NCS-1 age range of 15-54. The exclusion of the 15-17 age range was dictated by carrying out a separate NCS Adolescent survey of 10,000 respondents in the age range 13-17. The inclusion of the age range 55 years and older was based on the desire to study the entire adult age range. Part II was administered only to 5,692 of the 9,282 Part I respondents, including all Part I respondents with a lifetime disorder plus a probability subsample of other respondents.

- The details of the sample design are included in the paper

- Kessler, Ronald C.; Berglund, P.; Chiu, W.T.; Demler, O.; Heeringa, S.; Hiripi, E.; Jin, R.; Pennell, B.E.; Walters, E.E.; Zaslavsky, A.; Zheng, H., "The US National Comorbidity Survey Replication (NCS-R): Design and field procedures." *International Journal of Methods in Psychiatric Research*. 2004, 13, (2), 69 - 92.

NCS-R Sample Design

- Due the clustering of the NCS-R sample design, variance estimation from standard software procedures is incorrect
- NCS-R analysts should take the complex nature of the design into account by using SAS' surveyprocs, Sudaan, Stata's svy procs, or the Complex Samples module of SPSS
- Without doing the corrected variance/SE's the significance tests will be wrong, generally under-inflated
- Two key variables representing the clustering of the design are included in the NCS-R dataset: str (stratum) and SECU (Sampling Error Computing Unit)
- These corrections are included in all hands-on computer work

NCS-R Weights

- The NCS-R data are weighted to adjust for differential probabilities of selection of respondents within households and differential non-response as well as to adjust for residual differences between the sample and the United States population on the cross-classification of socio-demographic variables. An additional weight was used in the Part II sample to adjust for differences in probability of selection into that sample.
- These procedures are described in more detail by Kessler, Ronald C., Berglund, P., Chiu, W.T., et al., U.S. National Comorbidity Survey Replication (NCS-R): Design and Field Procedures, 2004.

Why Use Weights?

- Weighting is used to compensate for:
  - Unequal probabilities of selection
  - Nonresponse (typically, a unit that fails to respond)
  - In poststratification to adjust weighted sample distributions for certain variables (e.g., age and sex) to make them conform to the known population distribution.
  - It is used to improve the accuracy (minimize bias) of sample estimates and to compensate for noncoverage and nonresponse

Basic Weighting Approach

- Suppose sample element *i* was selected with probability p_i. Then sample element *i* represents ($1 / p_i$) elements in the population.
- That is, count the element i in the analysis by giving it a weight of
- $w_i = (1 / p_i)$.
- For example, a sample element selected with probability 1/10 represents 10 elements in the population.
- From Heeringa slides for Analysis of Complex Sample Survey Data

Overall Weight

- Weighting may incorporate simultaneously all three components, unequal probabilities of selection, nonresponse, and poststratification:
- Weight for unequal probabilities of selection: $w_{1;}$
- Weight for sample nonresponse: w_2;
- Poststratification weight for population noncoverage and sampling variance reduction: w_3.

Then compute the overall weight as:

$$w = w_1 \times w_2 \times w_3$$

Use of NCS-R Weights

NCS-R Weights:

- Part 1 weight sums to 9282
- Part 2 weight sums to 5692

General guidelines concerning which weight to use:

- If all variables in analysis come from the Part I of the instrument, use the Part 1 weight

- If you have either all Part 2 or a mix of Part 1 and Part 2 variables, use the Part 2 weight

Bias Example

Use of final weights is important to obtain correct, unbiased prevalences, as an example I present a table that outlines the effects of not using weights for Mexico, one of the countries in the WMH Initiative

Mexico
Table 1: Sociodemographic distribution of the Mexico sample compared to population[1]

	P1 Unweighted	P2 Unweighted	P1 Weighted	P2 Weighted	Census
Sex					
Male	39.4	36.1	47.6	47.7	47.7
Female	60.6	63.9	52.4	52.3	52.3
Age					
18-24	21.4	24.3	24.5	25.4	24.7
25-29	14.2	13.6	15.5	15.9	15.6
30-34	14.0	12.2	13.6	12.5	13.6
35-39	13.7	13.1	12.3	11.5	12.1
40-44	11.0	10.3	10.1	10.6	9.9
45-49	8.0	8.3	7.7	7.8	7.8
50-54	6.5	6.9	6.3	6.3	6.4
55-59	4.8	5.6	4.7	4.9	4.9
60-65	6.4	5.7	5.3	5.2	5.1
Region					
Metropolitan	31.5	32.6	28.0	27.6	27.6
Northwest	9.8	10.9	7.9	8.0	8.0
North	14.5	13.7	15.3	15.1	15.1
Central West	12.7	13.4	12.3	12.4	12.4
Central East	15.1	14.9	17.1	17.5	17.5
South East	16.4	14.6	19.4	19.3	19.3

[1]Presents only the sociodemographic variables used in post-stratification of weight.

Analysis of Complex Sample Data

- Data originates from sample designs that include features such as non-response adjustments, clustering, stratification, and differing probabilities of selection, NCS-R data is complex and adjustments are required for proper analysis

- Complex samples do not assume independence of observations, clustering and homogeneity are present

- Assuming a Simple Random Sample (SRS) generally results in underestimation of variance estimates due to effective loss of sample size due to clustering within strata

- Analysis of complex sample survey data should use variance estimation techniques that account for complex sample design features

- Software that uses correct analysis techniques are SAS surveyprocs, Sudaan, STATA svy procs, SPSS complex samples module, and other software that generally works under SAS (IVEware)

Design Variables

- Strata and Cluster variables are specified to represent the complex sample of the data, each country has the appropriate design variables in the demographic dataset

- From SAS v9.1.2 help
  - The STRATA statement names variables that form the strata in a stratified sample design. The combinations of categories of STRATA variables define the strata in the sample.

  - The CLUSTER statement names variables that identify the clusters in a clustered sample design. The combinations of categories of CLUSTER variables define the clusters in the sample. If there is a STRATA statement, clusters are nested within strata

NCS-R Design Variables

- Along with the correct weight, 2 key variables: str and secu
- Following, there is a cross tab of str*secu (stratum) and (Sampling Error Computing Unit)
- There are 42 stratum and values of 1 or 2 for each SECU in the NCS-R sample

CrossTab of str*secu variables

The FREQ Procedure

STR	SECU	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	41	0.44	41	0.44
1	2	50	0.54	91	0.98
2	1	59	0.64	150	1.62
2	2	49	0.53	199	2.14
3	1	68	0.73	267	2.88
3	2	55	0.59	322	3.47
4	1	68	0.73	390	4.20
4	2	66	0.71	456	4.91
5	1	63	0.68	519	5.59
5	2	61	0.66	580	6.25
6	1	57	0.61	637	6.86
6	2	56	0.60	693	7.47
7	1	92	0.99	785	8.46
7	2	96	1.03	881	9.49
8	1	66	0.71	947	10.20
8	2	91	0.98	1038	11.18
9	1	77	0.83	1115	12.01
9	2	63	0.68	1178	12.69
10	1	44	0.47	1222	13.17
10	2	82	0.88	1304	14.05
11	1	54	0.58	1358	14.63
11	2	53	0.57	1411	15.20
12	1	44	0.47	1455	15.68
12	2	64	0.69	1519	16.37
13	1	44	0.47	1563	16.84
13	2	54	0.58	1617	17.42
14	1	29	0.31	1646	17.73
14	2	50	0.54	1696	18.27
15	1	56	0.60	1752	18.88
15	2	53	0.57	1805	19.45
16	1	55	0.59	1860	20.04
16	2	52	0.56	1912	20.60
17	1	39	0.42	1951	21.02
17	2	35	0.38	1986	21.40
18	1	45	0.48	2031	21.88
18	2	41	0.44	2072	22.32
19	1	53	0.57	2125	22.89
19	2	60	0.65	2185	23.54
20	1	123	1.33	2308	24.87
20	2	137	1.48	2445	26.34
21	1	132	1.42	2577	27.76
21	2	177	1.91	2754	29.67

22	1	159	1.71	2913	31.38
22	2	138	1.49	3051	32.87
23	1	145	1.56	3196	34.43
23	2	124	1.34	3320	35.77
24	1	149	1.61	3469	37.37
24	2	137	1.48	3606	38.85
25	1	210	2.26	3816	41.11
25	2	168	1.81	3984	42.92
26	1	112	1.21	4096	44.13
26	2	160	1.72	4256	45.85
27	1	146	1.57	4402	47.43
27	2	133	1.43	4535	48.86
28	1	81	0.87	4616	49.73
28	2	133	1.43	4749	51.16
29	1	156	1.68	4905	52.84
29	2	180	1.94	5085	54.78
30	1	124	1.34	5209	56.12
30	2	115	1.24	5324	57.36
31	1	176	1.90	5500	59.25
31	2	217	2.34	5717	61.59
32	1	136	1.47	5853	63.06
32	2	153	1.65	6006	64.71
33	1	67	0.72	6073	65.43
33	2	82	0.88	6155	66.31
34	1	193	2.08	6348	68.39
34	2	168	1.81	6516	70.20
35	1	151	1.63	6667	71.83
35	2	192	2.07	6859	73.90
36	1	159	1.71	7018	75.61
36	2	146	1.57	7164	77.18
37	1	184	1.98	7348	79.16
37	2	149	1.61	7497	80.77
38	1	187	2.01	7684	82.78
38	2	179	1.93	7863	84.71
39	1	234	2.52	8097	87.23
39	2	167	1.80	8264	89.03
40	1	127	1.37	8391	90.40
40	2	186	2.00	8577	92.40
41	1	177	1.91	8754	94.31
41	2	169	1.82	8923	96.13
42	1	173	1.86	9096	98.00
42	2	186	2.00	9282	100.00

Common Statistical Techniques for Complex Sample Variance Estimation

- The Taylor Series Linearization Approach is used in all SAS SurveyProcs (surveymeans, surveyreg, surveyfreq, and surveylogistic) and Sudaan (also offers JRR)

- Other methods include Repeated Replication Techniques such as Jackknife Repeated Replication and Balanced Repeated Replication, these will not be demonstrated during this training session but are included in the references and can also be performed in Sudaan and STATA

- JRR is also easily programmed in SAS Macro language, see Berglund paper demonstrating the use of SAS macro for logistic regression JRR

- Use of design variables and weights will properly account for the complex sample structure

SAS v9.1

- SAS has full range of data management, analysis, and complex design correction procedures

- Complex design procs available
- Proc surveymeans-means, univariates
- Proc surveyfreq-frequency tables, 1 way and nway
- Proc surveyreg-linear dependent variables,ANOVA
- Proc surveylogistic-binary, ordinal, nominal logistic regression
- Proc surveyselect-sampling procedure

Domain and By-Group Analysis in SAS

- From SAS online help:
- The DOMAIN statement of SAS requests analysis for subpopulations, or domains, in addition to analysis for the entire study population. The DOMAIN statement names the variables that identify domains, which are called domain variables.

- It is common practice to compute statistics for domains. The formation of these domains may be unrelated to the sample design. Therefore, the sample sizes for the domains are random variables. In order to incorporate this variability into the variance estimation, you should use a DOMAIN statement.

- Note that a DOMAIN statement is different from a BY statement. In a BY statement, you treat the sample sizes as fixed in each subpopulation, and you perform analysis within each BY group independently

Sudaan/SAS/Stata Subgroup Analysis

- Sudaan software offers domain type analysis for every procedure via the "subpopn" statement
- SAS offers domain type analysis for descriptive procedures but not regression procs, Sudaan may offer a better choice for some types of regression analysis
- Use of a "subpopn" statement will allow Sudaan to subset the data for analysis but include all design information in the analysis, similar to the domain analysis in SAS
- Results from Sudaan with a subpopn statement will use the full degrees of freedom for the entire sample even though the actual analysis may not use all records
- Stata v9SE also offers a subpopulation analysis as well as a by group type of analysis, see the Stata documentation for help
- This techniques results in more degrees of freedom used in significance testing and leads to a conservative approach to testing

Planning an NCS-R Analysis

- Planning ahead always pays off when doing analysis, saves time and energy wasted on redoing variables or merging files

- Data management, variable creation or modification, descriptive exploration etc is generally a major part of doing good analysis, perhaps as high as 80-90% of analysis is preparing the data and examining results prior to modeling

- Subset the data by keeping only variables needed for analysis, this will greatly increase processing speed as the computer will not needlessly churn through thousands of unused variables

- Careful thought given to analysis methods used for type of research question: should it be means, frequencies, regression, survival analysis?

- How practical is it to use various small cells as predictors, do you have enough people with a condition of interest to do meaningful analysis?

Analysis Planning

- What software and hardware tools are needed to carry out analysis? Do you need to move to a remote server or can you do the work locally?
- What part of the NCS-R instrument do the variables of interest come from?
- What weights should be used?
- Has something very similar already been published by someone else? Is this a realistic topic?

Platform and Hardware Considerations

- **Personal Computer – Windows OS**
  - Compatible with other PC tools such as Office products, PC web-based tools, etc.
  - Generally not as fast for analysis and heavy computation as UNIX based systems
- **UNIX – Solaris/SUN OS or LINUX OS**
  - In general, better speed as compared to PC for complex analysis tasks
  - Better option if used as server for multiple users who will access datasets from central area and work simultaneously

Common Analysis Techniques

- Data analysis usually consists of 80-90% preparation for analysis by doing data management and data preparation, actual analysis phase is about 10-20% of task

- Of the 10-20% of analysis performed, about 80% of analysis would be covered by the following types of techniques:
- Descriptive analysis
  - Means/Univariates
  - Frequency tables
  - Graphs
  - Survival Curves

- Inferential analysis
  - Ordinary Least Squares
  - Logistic Regression with varying types of dependent variables (binary, ordinal, multinomial)
  - Survival analysis, mixed models/hierarchical models, Latent Class Analysis, and other more advanced methods represent a small portion of analyses

Standardized Approach to Analysis

- Standardized approach to analysis work
  - All team members use same software product, eliminates inefficiencies and confusion
  - Use coding rather than point and click, save programs for others to use, allows sharing of knowledge and programs
  - Replication of results can be achieved with organized setup and coding
  - Datasets stored in shared space and maintained by data manager
  - Shared computing resources, allows general sharing of programs and data, streamlined and less expensive licensing

Overview of Software : SAS v9.1

- SAS has full range of data management, analysis, graphing, reporting, complex design correction procedures and more

- Complex design procs available in v9.1 +
  - Proc surveymeans-means, univariates
  - Proc surveyfreq-frequency tables, 1 way and nway
  - Proc surveyreg-linear dependent variables,ANOVA
  - Proc surveylogistic-binary, ordinal, nominal logistic regression
  - Proc surveyselect-sampling procedure

SAS v9.1

- Compatible with external world, widely used in academia, government agencies, private business, etc.
- Main data management and analysis tool both Harvard Medical School and University of Michigan
- Offers ability to do entire range of tasks from data cleaning, preparation for analysis and complex design corrections, easy to use with very large files
- Compatible with other complex design sub-packages such as Sudaan, IVEware, and user-developed macros

SPSS v14

- SPSS is a very nice data management and basic analysis but no complex design corrections unless complex design module is purchased separately
- Complex design module available as a stand-alone module running with current Base version of SPSS
- Cost is quite high for complex design module, ($500 for just this module at University of Michigan) not cost-effective with small number of users
- Widely used around world, compatible with many users

STATA 9SE

- Stata offers full range of data management and analysis tools including many survey procedures for complex design corrections, offers varied options for complex design corrections
- Basically a stand-alone software, no other software runs under Stata as many do under SAS
- Not as widely used as SAS, more difficult to share code and datasets
- Not as good for large-scale data management tasks as compared to SAS
- All basic survey procs for descriptives and regression are included. Additional techniques are also available

Data Transfer Software

- DBMS Copy- allows easy and accurate movement of datafiles between all major packages such as SAS, SPSS, Excel, STATA
- StatTransfer-another good tool for moving data between software packages
- SAS built-in options- procs import and export, engine architecture allows reading of some types of external data sources, WIZARD enables point and click for data transfer
- SPSS- produces SAS files as output in SPSS14

Planning an NCS-R Analysis: Data Preparation

- Preparing data for analysis: missing data issues and common techniques for imputation: means, medians within subgroups, regression based imputation
- Recoding vars from 1 to 5 to 0/1, reversing scales, adding scales, arrays to process data iteratively, cleaning data by looking at outlier, wild codes, all missing right or wrong?
- Variable construction, example of complex variable such as time varying education or suicide ideation and onset derived from multiple questions

Data Preparation Techniques for Missing Data or Inconsistent Data

- Check for structural missing versus missing due to refusal or don't know by examining skip patterns and section flow
- Examine key variables needed for analysis to identify problems to fix or impute
- Examples are missing age of onset or age of recency along with disorder diagnosis
- Logical inconsistencies such as age of onset of disorder is later than age of first treatment for same disorder
- Check distributions for all variables needed for analysis to check for outliers or other problems prior to analysis phase

Strategies for dealing with missing data

- Imputation based on other variables that might give clues that will help assign a realistic imputed value, for example, one approach is to use age of first treatment or age of last episode to do best guess of age of onset if that is missing
- Use of overall statistic for a common crossing of demographic variables such as age*gender*education group that person falls within: assign mean or median for crosstab group for imputation of personal income
- Use regression based imputation tool such as IVEware or SAS Proc MI to impute values
- Check carefully for checkpoint problems or skip pattern inconsistency that might indicate data collection problem
- Check actual interview or respondent comments from text file to see if any further information available

Variable Recodes and Construction

- Amount of data management, variable preparation and recoding in typical NCS-R analysis is extensive and often quite complex, typically 80-90% of data analysis is preparing data for analysis

- Recodes can be simple things such as changing all 5's to 1's or reversing a scale, use of iterative coding such as arrays or macro do loops is an efficient way to handle this type of work

- Variable construction ranges from creating dummy variables to multi-layered variables that are created by complex recodes written within the macro language or other iterative processes

Example of Variable Construction

Ever talk to professional for particular disorder :

```
if d72 =1 then evertalkmde=1 ; else evertalkmde=0 ;
if d86 =1 then deptx12=1 ; else deptx12=0 ;
agetalkmde=d72a ;
if m33=1 then evertalkman=1 ; else evertalkman=0 ;
if m47=1 then mantx12=1 ; else mantx12=0 ;
agetalkman=m33a ;
```

repeat for all disorders used in lifetime treatment for disorder work

Example of More Complex Variable Construction with Macro Coding

```
*create variables that measure time between onset of dx and first tx for
    that dx* ;
%macro c (dx,suffix,onset) ;
*set people who talked to professional but wont give age to missing* ;
if agetalk&suffix ge 99 then agetalk&suffix=. ;
if &dx =1 and agetalk&suffix ne . then agetxint&suffix =agetalk&suffix ;
    else agetxint&suffix =age ;
    timeonsettx&suffix =(agetxint&suffix  - &onset ) ;
    ftimeonsettx&suffix =timeonsettx&suffix ;
    if timeonsettx&suffix ne . and timeonsettx&suffix lt 0 then do ;
    ftimeonsettx&suffix =0 ; agetxint&suffix=&onset  ;
    agetalk&suffix=&onset ;
end ;
%mend ;
%c (mde, mde, mde_ond) ;
```

Use of Array for Iterative Processing

```
array c1 [*] pt40b pt41b pt42b pt43b pt44b pt45b pt46b pt47b pt48b
   pt49b pt50b pt50_1b pt51b pt52b pt53b
   pt54b pt55b ;
   array c2 [*] ntillness ntbeatenup ntspouseabuse ntbeatnother ntmugged
   ntraped ntsexassualt ntstalked
   ntyoungdeath ntchildill nttrauma ntparentsfight ntdeadbody
   ntcausedeath ntkillother ntmasskilling
   ntotherevent  ;

   if private=1 then ntprivate=1 ; else ntprivate=. ;

   do i=1 to dim(c1) ;
        c2[i]=c1[i] ;
        if c1[i] in (995,998,999) then c2[i]=1 ;
        else if c1[i] > 10 then c2[i]=10 ;
   end ;
```

Scope of Variable Construction and Rules to Code By

- Most analyses involve hundreds of lines of coding to create, modify or collapse existing variables
- Types shown here are simply examples of the type of work done prior to descriptive or inferential analysis
- General rules are to create new variables when recoding, save all code so that results can be replicated at any point in time, write well-documented programs that can be understood by someone not familiar with project

Descriptive Analysis

- SAS SRS Procs: (produce Simple Random Sample statistics and variance estimates) : proc means, proc freq, proc univariate, proc corr, proc tabulate

- SAS Graphic and Reporting Procs: full range of graphing capacity in proc gplot, gchart, g3d. Reporting procs: proc report, proc print, proc tabulate

- SAS Complex sample procedures:
- means/corrected standard errors, difference in means
- proportions/corrected standard errors, chisq tests for tables

- SAS Survey Procs:
- proc surveymeans, proc surveyfreq

List of Contents of NCS-Public Release Dataset

- Top portion of contents output from dataset created from ASCII text file using SAS setup statements
- Note n=9282 and 4802 variables in file

```
                                  The CONTENTS Procedure

Data Set Name        D.ICPSRNCSR                     Observations          9282
Member Type          DATA                            Variables             4802
Engine               V9                              Indexes               0
Created              Tuesday, July 11,               Observation Length    38416
                     2006 12:46:59 PM
Last Modified        Tuesday, July 11,               Deleted Observations  0
                     2006 12:46:59 PM
Protection                                           Compressed            NO
Data Set Type                                        Sorted                NO
Label
Data Representation  WINDOWS_32
Encoding             wlatin1  Western (Windows)


                       Engine/Host Dependent Information

    Data Set Page Size          38912
    Number of Data Set Pages    9299
    First Data Page             18
    Max Obs per Page            1
    Obs in First Data Page      1
    Number of Data Set Repairs  0
    File Name                   f:\ncsr_training_july2006\icpsrncsr.sas7bdat
    Release Created             9.0101M3
    Host Created                XP_PRO
```

Contents Listing of NCS-R Dataset

SAS Code to produce contents listing:

```
options ls=90 ps=61 ;
 libname d 'f:\ncsr_training_july2006' ;
 libname library 'f:\ncsr_training_july2006' ;

*obtain contents of entire file first prior to subsetting* ;
 proc contents data=d.icpsrncsr ; run ;
```

Partial Output from Proc Contents:

```
Alphabetic List of Variables and Attributes

   #    Variable    Type    Len    Format     Label

1064    AAG3B1      Num       8    V571F.     AG3B: Approx age 1st fear alone/pub sit
1065    AAG3B2      Num       8    V31F.      AAG3B2: Before 1st started school, AG3B
1066    AAG3B3      Num       8    V31F.      AAG3B3: Before a teenager, AG3B
1067    AAG3B4      Num       8    V30F.      AAG3B4: Qualifier, AG3B
1076    AAG6A1      Num       8    V872F.     AG6A: Age 1st avd alon/pub sit bc fear
1077    AAG6A2      Num       8    V31F.      AAG6A2: Before 1st startd school, AAG6A1
1078    AAG6A3      Num       8    V31F.      AAG6A3: Before a teenager, AAG6A1
4147    AD3         Num       8    V4150F.    AD3: Mem ExctAge Vry1st DfcltCnc >=6mo
4154    AD4         Num       8    V4157F.    AD4: Still lot DfcltCnc/Atn drng payr
4156    AD5         Num       8    V226F.     AD5: #yrs altgthr have/hd DfcltCnc/Atn
4165    AD12        Num       8    V226F.     AD12: #d/365Payr TotUnablWrk/Act DfCnc
4166 AD14        Num       8    V4169F.    AD14: Ever talk MD/OthPro DfcltCnc/Atn
```

Examination of Analysis Variables

- The focus of the first part of the descriptive work will be an examination of the key demographic variables used in the NCS-R and selected key DSM disorders in the total sample and by gender

- Demographic variables studied:
  - age groups/cohorts (age at interview)
  - sex
  - race
  - marital status
  - region
  - education

- Selected Disorders
  - dsm_pds (DSM Panic Disorder)
  - dsm_so (social phobia)
  - dsm_sp (specific phobia)
  - dsm_gad (GAD)
  - dsm_ago (Agoraphobia)
  - dsm_pts (Post-Traumatic Stress Disorder)

SAS Code

Use of Proc Surveymeans for design corrected and weighted analysis of key demographic variables

```
data two  ;
    set one ;
*create a 4 category age variable for models * ;
if age <=29 then agecat=1 ;
else if 30<=age<=44 then agecat=2 ;
else if 45 <=age <=59 then agecat=3 ;
else if age >= 60 then agecat=4 ;

**create a dummy variable yes / no for ever thought about suicide* ;
if sd2 =1 or sd15=1 then suicideidea=1 ; else suicideidea=0 ;

*change value of negative 1 on part 2 weight to SAS system missing or . ;
if finalp2w eq -1 then finalp2w = . ;

*first step is to identify and examine each of the variables to be used in the analysis of anxiety
    disorders* ;
*proc contents ;
*run ;
proc format ;
value agef 1='<=29' 2='30-44' 3='45-59' 4='60+' ;

options ls=140 ps=49 orientation=landscape  ;
proc surveymeans mean stderr ;
strata str ;
cluster secu ;
weight finalp1w ;
class agecat educ_cat sex mar_stat region racecat_ ;
var agecat educ_cat sex mar_stat region racecat_ ;
format agecat agef. ;
run ;
```

SurveyMeans Analysis of Demographic Variables

Variable	Level	Label	Mean	Std Error of Mean
agecat	<=29		0.233557	0.012198
	30-44		0.295964	0.007323
	45-59		0.258883	0.007371
	60+		0.211595	0.006541
EDUC_CAT	(1) 0-11 YEARS EDUC	Education 4cat(NCS-R coding/non-imputed)	0.161366	0.006589
	(2) 12 YEARS EDUC	Education 4cat(NCS-R coding/non-imputed)	0.322478	0.012349
	(3) 13-15 YEARS EDUC	Education 4cat(NCS-R coding/non-imputed)	0.276669	0.007239
	(4) >=16 YEARS EDUC	Education 4cat(NCS-R coding/non-imputed)	0.239487	0.011605
SEX	(0) FEMALE	Sex	0.521126	0.005315
	(1) MALE	Sex	0.478874	0.005315
MAR_STAT	(1) MARRIED/COHABITATING	Marital category (imputed)	0.558366	0.011361
	(2) SEPARATED/WIDOWED/DIVORCED	Marital category (imputed)	0.204404	0.005973
	(3) NEVER MARRIED	Marital category (imputed)	0.237230	0.011353
REGION	(1) NORTHEAST	Region of country	0.192950	0.033446
	(2) MIDWEST	Region of country	0.231590	0.017551
	(3) SOUTH	Region of country	0.358342	0.020480
	(4) WEST	Region of country	0.217117	0.021724
RACECAT_	(1) HISPANIC	Race category (imputed)	0.108458	0.010253
	(2) BLACK	Race category (imputed)	0.115565	0.011185
	(3) OTHER	Race category (imputed)	0.043557	0.004085
	(4) WHITE	Race category (imputed)	0.732420	0.019470

SAS Code for Prevalence of Part 1 Disorders by Gender

```
proc surveymeans mean stderr  ;
options orientation=portrait ls=90 ps=61 ;
strata str ;
cluster secu ;
weight finalp1w ;
domain sex ;
var newpds newso newsp newgad newago ;
run ;
```

- Note that the part 1 weight is used for these disorders
- Also recall that the variables have been recoded to 0/1 rather than 1/5/missing in datastep above

Prevalence of Part 1 Disorders by Gender

The SURVEYMEANS Procedure

Data Summary

Number of Strata	42
Number of Clusters	84
Number of Observations	9282
Sum of Weights	9282.13

Statistics

Variable	Mean	Std Error of Mean
newpds	0.047038	0.002259
newso	0.120937	0.004036
newsp	0.125074	0.004059
newgad	0.077649	0.003371
newago	0.023884	0.001815

Domain Analysis: Sex

Sex	Variable	Mean	Std Error of Mean
(0) FEMALE	newpds	0.061962	0.003225
	newso	0.130395	0.006010
	newsp	0.157985	0.005804
	newgad	0.099472	0.004085
	newago	0.028765	0.002941
(1) MALE	newpds	0.030797	0.003344
	newso	0.110644	0.005677
	newsp	0.089258	0.005520
	newgad	0.053901	0.004690
	newago	0.018572	0.002476

SAS Surveymeans Output

- Analysis of the means by gender shows that women are much more likely to qualify for anxiety disorders
- This is also true for mood disorders but reverses for substance and impulse disorders
- With a domain statement you will receive the prevalences for the entire statement as well as for the values of the domain variable
- In this run, I have requested only 2 statistics but many more are available
- Another nice bit of output is the design variable summary area telling how many str and secu are present and the n of cases used
- The standard errors have been corrected taking the weights and the design variables into account and are in general, larger than those obtained from an SRS proc means analysis

SurveyMeans Analysis for Part 2 Disorder (PTSD)

```
proc surveymeans mean stderr nobs nmiss ;
title "Part 2 Disorder - Use finalp2w" ;
strata str ;
cluster secu ;
weight finalp2w ;
domain sex ;
var newpts ;
run ;
```

- Note use of finalp2w since PTSD is a part 2 disorder
- Same type of domain analysis, check output for different n (part2 n= 5692) instead of part 1 n=9282

Prevalence of PTSD by Gender

The SURVEYMEANS Procedure

Data Summary

Number of Strata	42
Number of Clusters	84
Number of Observations	9282
Number of Observations Used	5692
Number of Obs with Nonpositive Weights	3590
Sum of Weights	5692.29

Statistics

Variable	N	N Miss	Mean	Std Error of Mean
newpts	5692	0	0.068375	0.004272

Domain Analysis: Sex

Sex	Variable	N	N Miss	Mean	Std Error of Mean
(0) FEMALE	newpts	3310	0	0.097127	0.006607
(1) MALE	newpts	2382	0	0.035890	0.002927

Creating a Binary Outcome Variable: Suicide Ideation

- Suppose your analysis question is whether or not a person had a suicide ideation at some point in their lifetime
- Step 1 would be to read the entire suicide section of the instrument and plan out the construction of the suicide ideation variable
- After reading the instrument, it is clear that the ideation questions are in 2 places: one for those that can read and another spot for those that cannot read

This logic needs to be coded into the suicide ideation variables as follows:

```
if sd2 =1 or sd15=1 then suicideidea=1 ; else suicideidea=0 ;
```

- Note that this code sets suicideidea to 1 if a person answers yes on either SD2 or SD15
- Also note that this code sets everything not equal to 1 to 0, will include missing values, may not be an approach you want to use but in this case this provides a simple way to create a dummy variable

Check of Variable Construction

```
proc freq ;
options nodate nonumber ;
title "check suicide ideation variable construction" ;
     tables sd2 sd15 suicideidea ;
     weight finalp1w ;
     run ;
```

```
                              The FREQ Procedure

                         SD2: Evr seriously thght commit suicide

                                                                Cumulative    Cumulative
                               SD2    Frequency     Percent     Frequency      Percent
 ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
 REFUSED->GO TO SR1, NEXT SECTION         2.73        0.03           2.73          0.03
 DON'T KNOW->GO TO SR1, NEXT SECTION      7.53        0.08          10.26          0.11
 SYSTEM MISSING                         1634.3       17.61        1644.56         17.72
 (1) YES                               1204.69       12.98        2849.25         30.70
 (5) NO->GO TO SR1, NEXT SECTION       6432.88       69.30        9282.13        100.00


                        SD15: Evr seriously thght commit suicide

                                                                Cumulative    Cumulative
                              SD15    Frequency     Percent     Frequency      Percent
 ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
 REFUSED->GO TO SR1, NEXT SECTION         2.31        0.02           2.31          0.02
 DON'T KNOW->GO TO SR1, NEXT SECTION      1.82        0.02           4.13          0.04
 SYSTEM MISSING                        7646.77       82.38         7650.9         82.43
 (1) YES                                244.04        2.63        7894.94         85.06
 (5) NO->GO TO SR1, NEXT SECTION       1387.19       14.94        9282.13        100.00


                                                    Cumulative    Cumulative
             suicideidea    Frequency     Percent     Frequency      Percent
             ƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒƒ
                       0      7833.4       84.39        7833.4         84.39
                       1     1448.73       15.61       9282.13        100.00
```

Analysis of Suicide Ideation by Demographic Variables

```
***include suicide ideation as outcome*** ;
proc surveymeans mean stderr nobs nmiss ;
title "Analysis of Suicide Ideation among Demographic Subgroups" ;
strata str ;
cluster secu ;
weight finalp1w ;
domain agecat educ_cat sex mar_stat region racecat_ ;
class agecat educ_cat sex mar_stat region racecat_ ;
var suicideidea ;
run ;
```

- This code will give the prevalence of suicide ideation and corrected standard error for each category of the domain variables age, education, sex, marital status, region and race

Analysis of Suicide Ideation by Demo Variables

Analysis of Suicide Ideation among Demographic Subgroups

The SURVEYMEANS Procedure

Data Summary

Number of Strata	42
Number of Clusters	84
Number of Observations	9282
Sum of Weights	9282.13

Statistics

Variable	Mean	Std Error of Mean
suicideidea	0.156077	0.005176

Domain Analysis: agecat

agecat	Variable	Mean	Std Error of Mean
<=29	suicideidea	0.189929	0.010800
30-44	suicideidea	0.178707	0.009165
45-59	suicideidea	0.162585	0.007548
60+	suicideidea	0.079097	0.007327

Domain Analysis: Education 4cat(NCS-R coding/non-imputed)

Education 4cat(NCS-R coding/non-imputed)	Variable	Mean	Std Error of Mean
(1) 0-11 YEARS EDUC	suicideidea	0.190784	0.016575
(2) 12 YEARS EDUC	suicideidea	0.143348	0.008177
(3) 13-15 YEARS EDUC	suicideidea	0.158776	0.006776
(4) >=16 YEARS EDUC	suicideidea	0.146715	0.008414

Analysis of Suicide Ideation by Demo Variables

Domain Analysis: Sex

Sex	Variable	Mean	Std Error of Mean
(0) FEMALE	suicideidea	0.174115	0.005856
(1) MALE	suicideidea	0.136449	0.006039

Domain Analysis: Marital category (imputed)

Marital category (imputed)	Variable	Mean	Std Error of Mean
(1) MARRIED/COHABITATING	suicideidea	0.129053	0.005222
(2) SEPARATED/WIDOWED/DIVORCED	suicideidea	0.176909	0.008930
(3) NEVER MARRIED	suicideidea	0.201735	0.012833

Domain Analysis: Region of country

Region of country	Variable	Mean	Std Error of Mean
(1) NORTHEAST	suicideidea	0.145311	0.013136
(2) MIDWEST	suicideidea	0.153239	0.010310
(3) SOUTH	suicideidea	0.136869	0.007041
(4) WEST	suicideidea	0.200376	0.010074

Domain Analysis: Race category (imputed)

Race category (imputed)	Variable	Mean	Std Error of Mean
(1) HISPANIC	suicideidea	0.151671	0.014719
(2) BLACK	suicideidea	0.136414	0.011606
(3) OTHER	suicideidea	0.199283	0.023124
(4) WHITE	suicideidea	0.157263	0.005479

Descriptive Analysis Summary

- Plan the analysis using NCS-R data
- Select variables to be used from the correct section of the instrument and identify the correct sample weight
- Perform design-corrected and weighted means and standard errors for prevalences
- Create outcome variable for use in either descriptive or inferential analysis
- Do “domain” or “sub-population” analysis in SAS by using the domain statement in PROC SURVEYMEANS
- The next section of the training will focus on using the suicide ideation outcome as a binary dependent variable in a logistic regression and use SAS PROC SURVEYLOGISTIC for corrected SE’s

Hands-On Exercises

1. Explore the ICPSR and NCS websites and try to become comfortable with obtaining information you will need when you return to your homes.

Replication of the descriptive analyses presented

2. The exercises are pre-programmed for those of you that are not SAS programmers. On your CD you will find a SAS program called NCSRtraining.sas. Please open that program in the SAS program editor and replicate the output that we have just covered.

For those of you that already know SAS, try replicating everything we have covered thus far but write your own code instead of using ours. Feel free to use our code as a guideline.

Mike Gruber and I will come around and help with individual problems.

Logistic Regression

- For logistic regression we present an example of life-time suicide ideation predicted by common demographic covariates plus the anxiety disorders introduced in the previous section of the training
- This is a typical approach of using logistic regression with design-corrected standard errors and confidence intervals to predict a binary outcome

Sas Code for Logistic Regression

```
***logistic regression with suicide ideation as dependent variable,
 predictors are all disorders and demo variables * ;

proc surveylogistic ;
strata str ;
cluster secu ;
weight finalp2w ;
class agecat educ_cat sex mar_stat region racecat_  / param=reference ;
model suicideidea (event='1') = agecat educ_cat sex mar_stat region
   racecat_ newpds newso newsp newgad newago newpts ;
title "Suicide Ideation during Lifetime predicted by demographic
   variables and anxiety disorders" ;
run ;
```

- Note that SURVEYLOGISTIC uses the same type of specification of design variables and weight, why use of part 2 weight in this analysis?
- Code directly specifies that the outcome ‘event=1’ meaning the probability of suicide ideation is being predicted
- Use of class statement with /param=reference changes from the default of effect coding with a class statement in SAS SURVEYLOGISTIC

Logistic Results

Suicide Ideation during Lifetime predicted by demographic variables and anxiety disorders

The SURVEYLOGISTIC Procedure

Model Information

Data Set	WORK.TWO	
Response Variable	suicideidea	
Number of Response Levels	2	
Stratum Variable	STR	Strata NCS-R version
Number of Strata	42	
Cluster Variable	SECU	Sampling error computation unit
Number of Clusters	84	
Weight Variable	FINALP2W	Final part 2 weight
Model	Binary Logit	
Optimization Technique	Fisher's Scoring	
Variance Adjustment	Degrees of Freedom (DF)	

Number of Observations Read	9282
Number of Observations Used	5692
Sum of Weights Read	5692.29
Sum of Weights Used	5692.29

Response Profile

Ordered Value	suicideidea	Total Frequency	Total Weight
1	0	4347	4805.2200
2	1	1345	887.0700

Probability modeled is suicideidea=1.

NOTE: 3590 observations having nonpositive frequencies or weights were excluded since they do not contribute to the analysis.

Results, continued

Class Level Information

Class	Value	Design Variables		
agecat	1	1	0	0
	2	0	1	0
	3	0	0	1
	4	0	0	0
EDUC_CAT	(1) 0-11 YEARS EDUC	1	0	0
	(2) 12 YEARS EDUC	0	1	0
	(3) 13-15 YEARS EDUC	0	0	1
	(4) >=16 YEARS EDUC	0	0	0
SEX	(0) FEMALE	1		
	(1) MALE	0		
MAR_STAT	(1) MARRIED/COHABITATING	1	0	
	(2) SEPARATED/WIDOWED/DIVORCED	0	1	
	(3) NEVER MARRIED	0	0	

Class Level Information

Class	Value	Design Variables		
REGION	(1) NORTHEAST	1	0	0
	(2) MIDWEST	0	1	0
	(3) SOUTH	0	0	1
	(4) WEST	0	0	0
RACECAT_	(1) HISPANIC	1	0	0
	(2) BLACK	0	1	0
	(3) OTHER	0	0	1
	(4) WHITE	0	0	0

Results, continued

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	4928.130	4382.646
SC	4934.776	4528.876
-2 Log L	4926.130	4338.646

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	587.4841	21	<.0001
Score	690.9353	21	<.0001
Wald	785.8907	21	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
agecat	3	60.4565	<.0001
EDUC_CAT	3	16.2470	0.0010
SEX	1	4.9035	0.0268
MAR_STAT	2	47.0164	<.0001
REGION	3	10.7966	0.0129
RACECAT_	3	17.2512	0.0006
newpds	1	15.4733	<.0001
newso	1	63.0151	<.0001
newsp	1	5.0391	0.0248
newgad	1	57.3722	<.0001
newago	1	2.3079	0.1287

Results, continued

The SURVEYLOGISTIC Procedure

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
newpts	1	79.6897	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.4234	0.1845	172.5490	<.0001
agecat	1	1	0.9021	0.1418	40.4478	<.0001
agecat	2	1	0.9470	0.1246	57.7734	<.0001
agecat	3	1	0.7904	0.1335	35.0709	<.0001
EDUC_CAT	(1) 0-11 YEARS EDUC	1	0.5184	0.1419	13.3373	0.0003
EDUC_CAT	(2) 12 YEARS EDUC	1	0.0150	0.1098	0.0186	0.8916
EDUC_CAT	(3) 13-15 YEARS EDUC	1	0.0530	0.0909	0.3402	0.5597
SEX	(0) FEMALE	1	0.1343	0.0606	4.9035	0.0268
MAR_STAT	(1) MARRIED/COHABITATING	1	-0.5352	0.1120	22.8415	<.0001
MAR_STAT	(2) SEPARATED/WIDOWED/DIVORCED	1	-0.1305	0.1415	0.8510	0.3563
REGION	(1) NORTHEAST	1	-0.2944	0.1503	3.8371	0.0501
REGION	(2) MIDWEST	1	-0.2631	0.1079	5.9487	0.0147
REGION	(3) SOUTH	1	-0.3463	0.1117	9.6182	0.0019
RACECAT_	(1) HISPANIC	1	-0.3590	0.1334	7.2441	0.0071
RACECAT_	(2) BLACK	1	-0.4407	0.1245	12.5401	0.0004
RACECAT_	(3) OTHER	1	0.0495	0.1624	0.0930	0.7604
newpds		1	0.6325	0.1608	15.4733	<.0001
newso		1	0.8085	0.1018	63.0151	<.0001
newsp		1	0.2424	0.1080	5.0391	0.0248
newgad		1	0.9395	0.1240	57.3722	<.0001
newago		1	0.2812	0.1851	2.3079	0.1287
newpts		1	0.9469	0.1061	79.6897	<.0001

Results, continued

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
agecat 1 vs 4	2.465	1.867	3.255
agecat 2 vs 4	2.578	2.019	3.291
agecat 3 vs 4	2.204	1.697	2.863
EDUC_CAT (1) 0-11 YEARS EDUC vs (4) >=16 YEARS EDUC	1.679	1.271	2.218
EDUC_CAT (2) 12 YEARS EDUC vs (4) >=16 YEARS EDUC	1.015	0.819	1.259
EDUC_CAT (3) 13-15 YEARS EDUC vs (4) >=16 YEARS EDUC	1.054	0.882	1.260
SEX (0) FEMALE vs (1) MALE	1.144	1.016	1.288
MAR_STAT (1) MARRIED/COHABITATING vs (3) NEVER MARRIED	0.586	0.470	0.729
MAR_STAT (2) SEPARATED/WIDOWED/DIVORCED vs (3) NEVER MARRIED	0.878	0.665	1.158
REGION (1) NORTHEAST vs (4) WEST	0.745	0.555	1.000
REGION (2) MIDWEST vs (4) WEST	0.769	0.622	0.950
REGION (3) SOUTH vs (4) WEST	0.707	0.568	0.880
RACECAT_ (1) HISPANIC vs (4) WHITE	0.698	0.538	0.907
RACECAT_ (2) BLACK vs (4) WHITE	0.644	0.504	0.821
RACECAT_ (3) OTHER vs (4) WHITE	1.051	0.764	1.444
newpds	1.882	1.373	2.580
newso	2.244	1.838	2.740
newsp	1.274	1.031	1.575
newgad	2.559	2.007	3.263
newago	1.325	0.922	1.904
newpts	2.578	2.094	3.173

Association of Predicted Probabilities and Observed Responses

Percent Concordant	67.8	Somers' D	0.361
Percent Discordant	31.6	Gamma	0.364
Percent Tied	0.6	Tau-a	0.130
Pairs	5846715	c	0.681

Interpretation of Logistic Regression Output

- Probability modeled is suicideidea=1
- Because of the use of the part 2 weight, n=5692
- The overall model fit is significant with a Wald Chi-Square of 785.89 with 21 df's and a p value of <.0001
- The Wald ChiSq tests for the predictor variables show that all predictors are significant with the exception of agoraphobia
- The Odds Ratios show that
  - younger age groups are more likely to have suicide ideation, as compared to the oldest age group
  - the least educated group is more likely to have suicide ideation, as compared to the highly educated (16+)
  - women are slightly more likely than men to have suicide ideation
  - those that live in the West region are more likely than those from all other regions to have suicide ideation
  - hispanics and blacks are significantly less likely than whites to have suicide ideation
  - with the exception of those with agoraphobia, having any of the anxiety disorders makes you more likely than those with no anxiety disorder to have suicide ideation

Adding Linear Contrasts to Logistic Regression

- Suppose you want to add a test of significance between race categories
- This examples illustrates how this can be done using the contrast statement in SAS SURVEYLOGISTIC

```
proc surveylogistic ;
strata str ;
cluster secu ;
weight finalp2w ;
class agecat educ_cat sex mar_stat region racecat_  /
   param=reference ;
model suicideidea (event='1') = agecat educ_cat sex mar_stat
   region racecat_ newpds newso newsp newgad newago newpts ;
contrast "Test Hispanics versus Blacks" racecat_ 1 -1 0 ;
contrast "Test Blacks versus Other" racecat_ 0 1 -1 ;
title "Suicide Ideation during Lifetime predicted by
   demographic variables and anxiety disorders" ;
run ;
```

Contrasts

- Partial Output from SURVEYLOGISTIC
- This output shows that there is a significant difference between Blacks and Other but not between Hispanics and Blacks

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
Test Hispanics versus Blacks	1	0.2414	0.6232
Test Blacks versus Other	1	5.0541	0.0246

Subpopulation Analyses in SAS SurveyProcedures

- As of SAS v9.1.3 neither the SURVEYREG or the SURVEYLOGISTIC procedures contain a “domain” statement
- This can potentially present problems if you want to analyze small subpopulations where the possibility of zero cells in the str*secu matrix occur
- NCS-R analysts often do analyses such as these but use Sudaan (SAS-callable) with a subpopn statement
- The Sudaan software always uses the full design variable matrix even for subpopulation analyses and does not fail to run as SAS sometimes will
- It is beyond the scope of this training to demonstrate Sudaan but please be aware of this issue when analyzing subpopulations
- Other software choices such as IVEware, Stata, SPSS complex samples will have similar considerations

Linear Regression

- A majority of NCS-R analyses are performed using logistic regression, due to the nature of many of our variables, many are categorical

- Linear regression is another important tool used in the analysis of NCS-R data but time constraints limit what we can present in this analysis

- Use of SAS PROC SURVEYREG or Sudaan proc regress or a similar design corrected linear regression tool is recommended

- Please see software documentation for help and examples

Computer Exercises for Logistic Regression

- Using the same SAS/non-SAS user strategy (use our code or write your own), replicate the logistic regression demonstrated
- If you want a challenge try using other DSM disorders than what I have presented. Good choices might be substance disorders or mood disorders. (Alcohol abuse/dependence, drug abuse/dependence, MDDH or MDE, bipolar disorders, dysthymia)
- Another challenge: try using PROC SURVEYREG with a linear outcome variable such as household income

Survival Analysis

- Timing of events is studied in event history analysis, analysis of record of when events occur

- Predictors can be either time-varying (marital status, education) or time-invariant (race)

- Censoring is another key concept for event history analysis, censoring occurs when follow-up of individual ends and we can no longer determine whether or not event of interest occurs, such cases that do not yet have the outcome of interest are called censored

- Right censoring occurs naturally in our surveys, censored at time of interview

Discrete-Time and Continuous Time

- Continuous time is measured as a positive, continuous variable
- Discrete-Time is appropriate for situations in which events can only occur at regular point in time: presidential elections every four years, yearly medical examination, yearly inteview
- Discrete-time can also handle situations where an event can occur at any point in time yet data is available or collected at a certain discrete point in time, our surveys typically deal with data of this type that asks only if a marital status change occurred during this year or did you have onset of a disorder in this year
- Most survival analyses with NCS-R related datasets are done using the discrete-time approach

Data Structure and Key Variables

- For survival analysis you need a few key variables:
  - Time at which event of interest occurred and time at which last observed if event did not occur, example is age of onset of a given disorder or age at interview if no disorder
  - Status of event occurring or yes/no type variable, for example dummy variable for disorder 0=no, 1=yes
  - With time and status or age at onset or age at interview and yes/no for disorder we now have the key variables to use in survival curve analysis

Age of Onset of Disorder Survival Curve

- A common approach is to use an age of onset for a given disorder or group of disorders and graph the cumulative frequencies or percentages for the onset ages
- Key variables in this type of analysis are yes/no indicator of having the disorder of interest, age at onset of the disorder, age censored or usually age of interview for our needs
- Data is structured with 1 record per person and proc lifetest of SAS is used to examine the survival/failure distribution
- Use of survival curves is intended to be descriptive or exploratory in nature in the following examples

Sample data for survival curve

- Sample records with key variables detailed:

Sampleid	dsm_mde	mde_ond	age
1	0	0	50
2	1	20	44
etc.			

Person number 1 would be followed from years 1 to 50 and would never have a “yes” on the outcome of dsm_mde

- Person number 2 would be followed from years 1 to 20 (age of onset of dsm_mde) and has a “yes” on dsm_mde as well as an age of onset for mde

- Note that every person will be included in this type of analysis since we have year 1 and age at interview for every respondent

Survival Curve for Major Depressive Disorder

- Outcome variable is dsm_mde (major depressive disorder)
- Age of onset of disorder is mde_ond
- Age of interview is used as the censor variable if no disorder present

Overall Prevalence for Major Depressive Disorder and Age of Onset Distribution

```
proc surveymeans data=two  ;
strata str ;
cluser secu ;
weight finalp1w ;
var dsm_mde ;
run ;

proc freq data=two  ;
tables mde_ond* dsm_mde;
weight finalp1w ;
run ;
```

SURVEYMEANS Output for MDE

```
The SURVEYMEANS Procedure

                                   Data Summary

                        Number of Strata                      42
                        Number of Clusters                    84
                        Number of Observations              9282
                        Sum of Weights                   9282.13


                                    Statistics

  Variable    Label                                                N            Mean
  ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
  DSM_MDE     DSM-IV MajorDepressiveEpisode(Lifetime)           9282        0.191697
  ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff

                                    Statistics

                               Std Error
                   Variable      of Mean       95% CL for Mean
                   fffffffffffffffffffffffffffffffffffffffffffffffff
                   DSM_MDE      0.004877    0.18185583 0.20153893
                   fffffffffffffffffffffffffffffffffffffffffffffffff
```

The FREQ Procedure

MDE_OND	DSM_MDE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
SYSTEM MISSING	0	7502.77	80.83	7502.77	80.83
4	(1) ENDORSED	17.47	0.19	7520.24	81.02
5	(1) ENDORSED	17.58	0.19	7537.82	81.21
6	(1) ENDORSED	20.33	0.22	7558.15	81.43
7	(1) ENDORSED	17.7	0.19	7575.85	81.62
8	(1) ENDORSED	21.97	0.24	7597.82	81.85
9	(1) ENDORSED	15.13	0.16	7612.95	82.02
10	(1) ENDORSED	28.82	0.31	7641.77	82.33
11	(1) ENDORSED	25.64	0.28	7667.41	82.60
12	(1) ENDORSED	76.22	0.82	7743.63	83.43
13	(1) ENDORSED	74.11	0.80	7817.74	84.22
14	(1) ENDORSED	52.85	0.57	7870.59	84.79
15	(1) ENDORSED	67.93	0.73	7938.52	85.52
16	(1) ENDORSED	86.72	0.93	8025.24	86.46
17	(1) ENDORSED	56.43	0.61	8081.67	87.07
18	(1) ENDORSED	68.12	0.73	8149.79	87.80
19	(1) ENDORSED	67.29	0.72	8217.08	88.53
20	(1) ENDORSED	56.85	0.61	8273.93	89.14
21	(1) ENDORSED	41.83	0.45	8315.76	89.59
22	(1) ENDORSED	47.98	0.52	8363.74	90.11
23	(1) ENDORSED	42.71	0.46	8406.45	90.57
24	(1) ENDORSED	39.54	0.43	8445.99	90.99
25	(1) ENDORSED	58.7	0.63	8504.69	91.62
26	(1) ENDORSED	24.27	0.26	8528.96	91.89
27	(1) ENDORSED	37.7	0.41	8566.66	92.29
28	(1) ENDORSED	36.23	0.39	8602.89	92.68
29	(1) ENDORSED	25.39	0.27	8628.28	92.96
30	(1) ENDORSED	47.12	0.51	8675.4	93.46
31	(1) ENDORSED	27.42	0.30	8702.82	93.76
32	(1) ENDORSED	45.8	0.49	8748.62	94.25
33	(1) ENDORSED	18.46	0.20	8767.08	94.45
34	(1) ENDORSED	41.62	0.45	8808.7	94.90
35	(1) ENDORSED	46.33	0.50	8855.03	95.40
36	(1) ENDORSED	22.64	0.24	8877.67	95.64
37	(1) ENDORSED	33.59	0.36	8911.26	96.00
38	(1) ENDORSED	30.66	0.33	8941.92	96.33
39	(1) ENDORSED	23.25	0.25	8965.17	96.59
40	(1) ENDORSED	37.25	0.40	9002.42	96.99
41	(1) ENDORSED	20.4	0.22	9022.82	97.21
42	(1) ENDORSED	35.8	0.39	9058.62	97.59
43	(1) ENDORSED	19.89	0.21	9078.51	97.81
44	(1) ENDORSED	10.3	0.11	9088.81	97.92
45	(1) ENDORSED	24.64	0.27	9113.45	98.18
46	(1) ENDORSED	14.48	0.16	9127.93	98.34
47	(1) ENDORSED	11.07	0.12	9139	98.46
48	(1) ENDORSED	18.39	0.20	9157.39	98.66

49	(1) ENDORSED	18.02	0.19	9175.41	98.85
50	(1) ENDORSED	13.72	0.15	9189.13	99.00
51	(1) ENDORSED	10.34	0.11	9199.47	99.11
52	(1) ENDORSED	5.3	0.06	9204.77	99.17
53	(1) ENDORSED	11.27	0.12	9216.04	99.29
54	(1) ENDORSED	5.58	0.06	9221.62	99.35
55	(1) ENDORSED	7.68	0.08	9229.3	99.43
56	(1) ENDORSED	4.33	0.05	9233.63	99.48
57	(1) ENDORSED	5.21	0.06	9238.84	99.53
58	(1) ENDORSED	2.16	0.02	9241	99.56
59	(1) ENDORSED	3.5	0.04	9244.5	99.59
60	(1) ENDORSED	3.7	0.04	9248.2	99.63
61	(1) ENDORSED	2.12	0.02	9250.32	99.66
62	(1) ENDORSED	1.96	0.02	9252.28	99.68
63	(1) ENDORSED	3.85	0.04	9256.13	99.72
64	(1) ENDORSED	1.65	0.02	9257.78	99.74
65	(1) ENDORSED	3.74	0.04	9261.52	99.78
66	(1) ENDORSED	1.87	0.02	9263.39	99.80
67	(1) ENDORSED	0.59	0.01	9263.98	99.80
68	(1) ENDORSED	2.21	0.02	9266.19	99.83
69	(1) ENDORSED	1.8	0.02	9267.99	99.85
70	(1) ENDORSED	2.47	0.03	9270.46	99.87
72	(1) ENDORSED	0.99	0.01	9271.45	99.88
74	(1) ENDORSED	1.18	0.01	9272.63	99.90
76	(1) ENDORSED	0.77	0.01	9273.4	99.91
78	(1) ENDORSED	0.94	0.01	9274.34	99.92
80	(1) ENDORSED	3.25	0.04	9277.59	99.95
81	(1) ENDORSED	0.58	0.01	9278.17	99.96
83	(1) ENDORSED	3.38	0.04	9281.55	99.99
86	(1) ENDORSED	0.58	0.01	9282.13	100.00

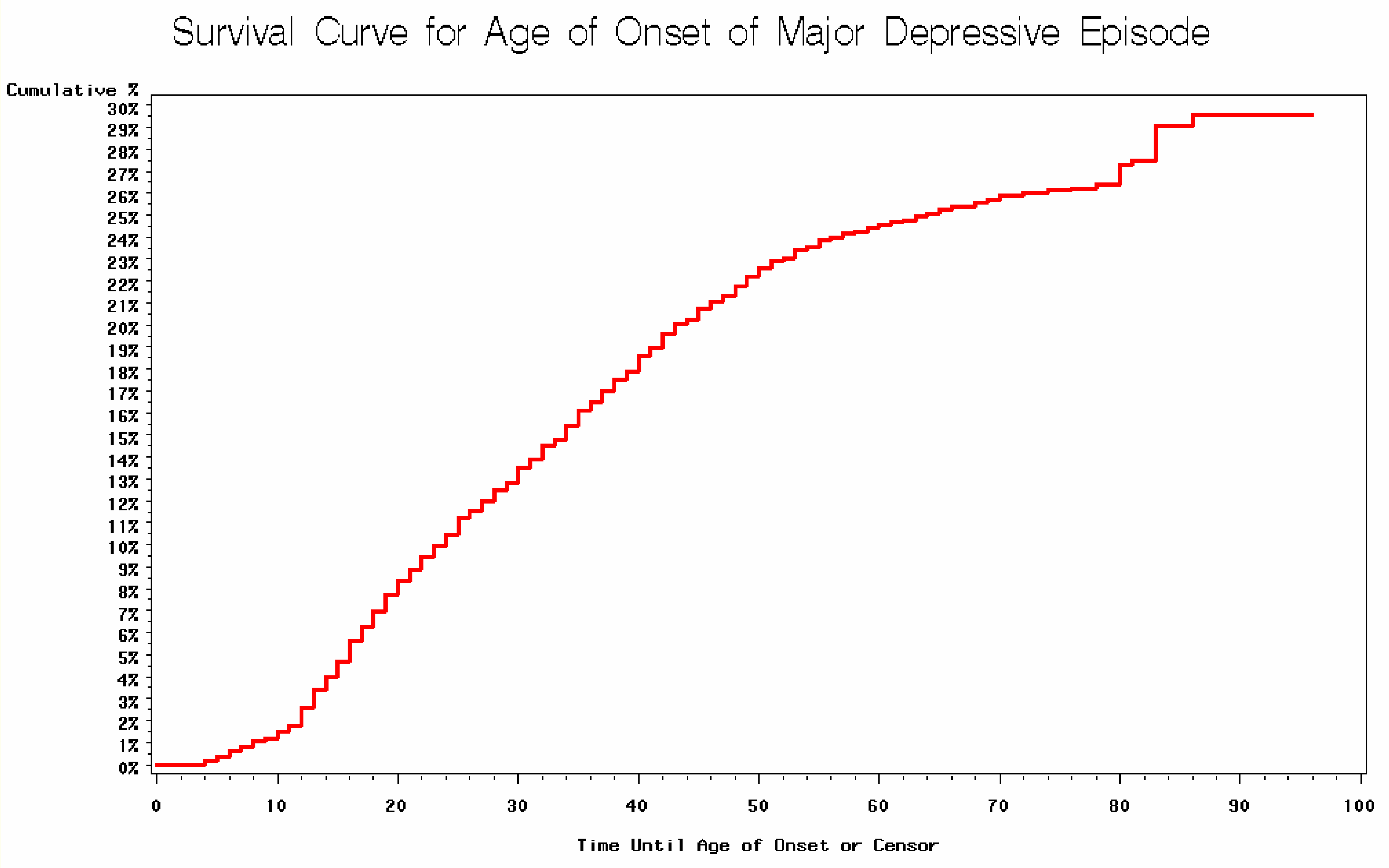
Survival Curve for Major Depressive Disorder

```
**prepare mde for survival curve analysis* ;
*recode dsm_mde* ;
if dsm_mde ne 1 then dsm_mde=0 ;
*create age at onset or age at censor* ;
if dsm_mde=1 then ageevent=mde_ond ; else ageevent=age ;

*multiply weight by 100 for proc lifetest freq statement ;
finalp1w100=finalp1w*100 ;

proc lifetest method=lt intervals=(1 to 96 by 1) notable outs=out  ;
    time ageevent * dsm_mde (0) ;
    freq finalp1w100 ;
run ;
data survival  ;
    set out ;
    fail=1-survival ;
label fail="Cumulative %" ageevent="Time Until Age of Onset or Censor" ;
proc print ;
run ;
symbol c=red i=steprj w=3 ;
title "Survival Curve for Age of Onset of Major Depressive Episode" ;
proc gplot ;
          plot fail*ageevent / legend  ;
          format fail percent10. ;
          run ;
```

MDE Survival Curve



Output from Proc Lifetest

Life Table Survival Estimates

Interval [Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime
0	1	0	0	923589.0	0	0	1.0000	0	0	.
1	2	0	0	923589.0	0	0	1.0000	0	0	.
2	3	0	0	923589.0	0	0	1.0000	0	0	.
3	4	0	0	923589.0	0	0	1.0000	0	0	.
4	5	1738	0	923589.0	0.00188	0.000045	1.0000	0	0	.
5	6	1748	0	921851.0	0.00190	0.000045	0.9981	0.00188	0.000045	.
6	7	2018	0	920103.0	0.00219	0.000049	0.9962	0.00377	0.000064	.
7	8	1759	0	918085.0	0.00192	0.000046	0.9940	0.00596	0.000080	.
8	9	2186	0	916326.0	0.00239	0.000051	0.9921	0.00786	0.000092	.
9	10	1504	0	914140.0	0.00165	0.000042	0.9898	0.0102	0.000105	.
10	11	2868	0	912636.0	0.00314	0.000059	0.9881	0.0119	0.000113	.
11	12	2550	0	909768.0	0.00280	0.000055	0.9850	0.0150	0.000126	.
12	13	7577	0	907218.0	0.00835	0.000096	0.9823	0.0177	0.000137	.
13	14	7376	0	899641.0	0.00820	0.000095	0.9741	0.0259	0.000165	.
14	15	5257	0	892265.0	0.00589	0.000081	0.9661	0.0339	0.000188	.
15	16	6758	0	887008.0	0.00762	0.000092	0.9604	0.0396	0.000203	

Survival and Failure Calculations

- Lifetable approach assumes that any censored cases will be censored in the midpoint of the time interval

- Lifetable approach is appropriate for situations with a large number of observations and many unique event times

- Survival is calculated from the conditional probabilities of failure as follows:
- Let: ti = time interval start time
- qi= conditional probability of failure

- The probability of surviving to t(i) or beyond is calculated as a series of probabilities:
- For example for t4:
- a=survival to t2 or beyond
- b=survival to t3 or beyond
- c=survival to t4 or beyond
- So, prC=pr(A, B, C) or
- pr(C)=(1-q3)(1-q2)(1-q1)
- Failure is the key statistic we graph: failure=1-survival (calculated from the survival formula above), Y axis is then cumulative % rather than regular % of a fixed denominator

Example of a More Complex Curve

- This graph examines time between year 1 of life and onset of individual anxiety disorders of social phobia, and specific phobia
- Variables are organized to measure time between year 1 of life and either onset of use or age at interview (censor)
- This approach again uses the discrete-time concept with proc lifetest
- Next we group all anxiety curves and examine in one graph
- Use of SAS ODS tools and graphing tools (gplot) are demonstrated

SAS Proc Lifetest and Graphing Code

```
proc lifetest data=two  method=lt intervals=(1 to 96 by 1) notable
   outs=outsp  ;
   time ageeventsp * newsp (0) ;
   freq finalp1w100 ;
run ;

data outspf ;
    set outsp ;
    fail=1-survival ;
label fail="Cumulative %" ageeventsp ="Time Until Age of Onset or
   Censor" ;
run ;

proc lifetest data=two method=lt intervals=(1 to 96 by 1) notable
   outs=outso  ;
   time ageeventso * newso (0) ;
   freq finalp1w100 ;
run ;

data outsof  ;
    set outso ;
    fail=1-survival ;
label fail="Cumulative %" ageeventso ="Time Until Age of Onset or
   Censor" ;
run ;
```

SAS Proc Lifetest and Graphing Code, continued

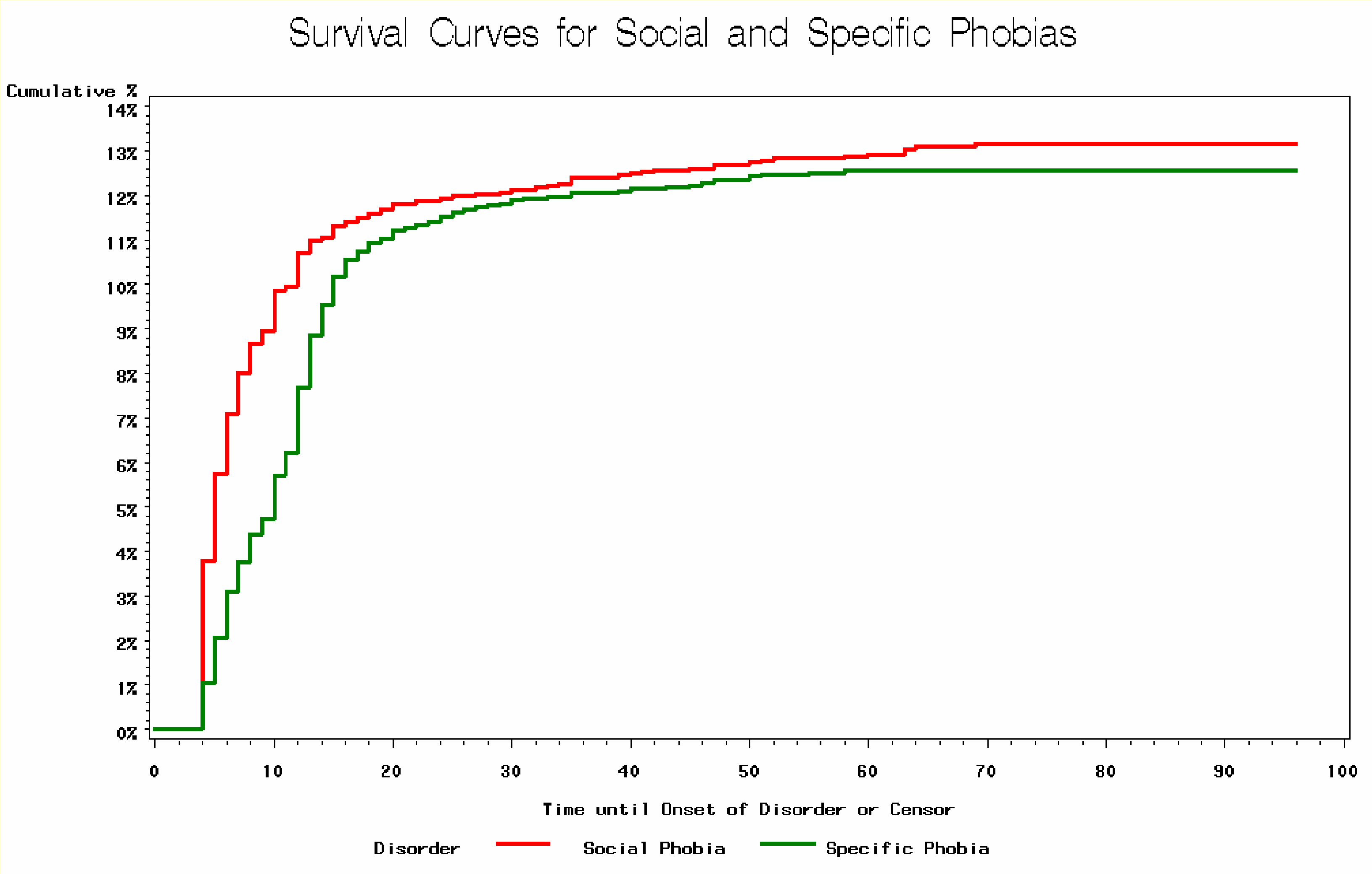
```
data allanx ;
set outspf (in=sp) outsof (in=so) ;
    if sp then type=1 ;
    if so then type=2 ;

    if type=1 then do ; timeonset=ageeventsp ;  end ;
    if type=2 then do ; timeonset=ageeventso   ; end ;
label timeonset="Time until Onset of Disorder or Censor" type="Disorder" ;
symbol c=red i=steprj w=3 ;
symbol2 c=green i=steprj w=3 ;
title "Survival Curve for Age of Onset of Major Depressive Episode" ;

proc format ;
value typef 1="Social Phobia" 2="Specific Phobia" ;

proc gplot ;
title "Survival Curves for Social and Specific Phobias" ;
          plot fail*timeonset=type  / legend  ;
          format fail percent10. type typef.  ;
          run ;
```

Survival Curves for Two Disorders



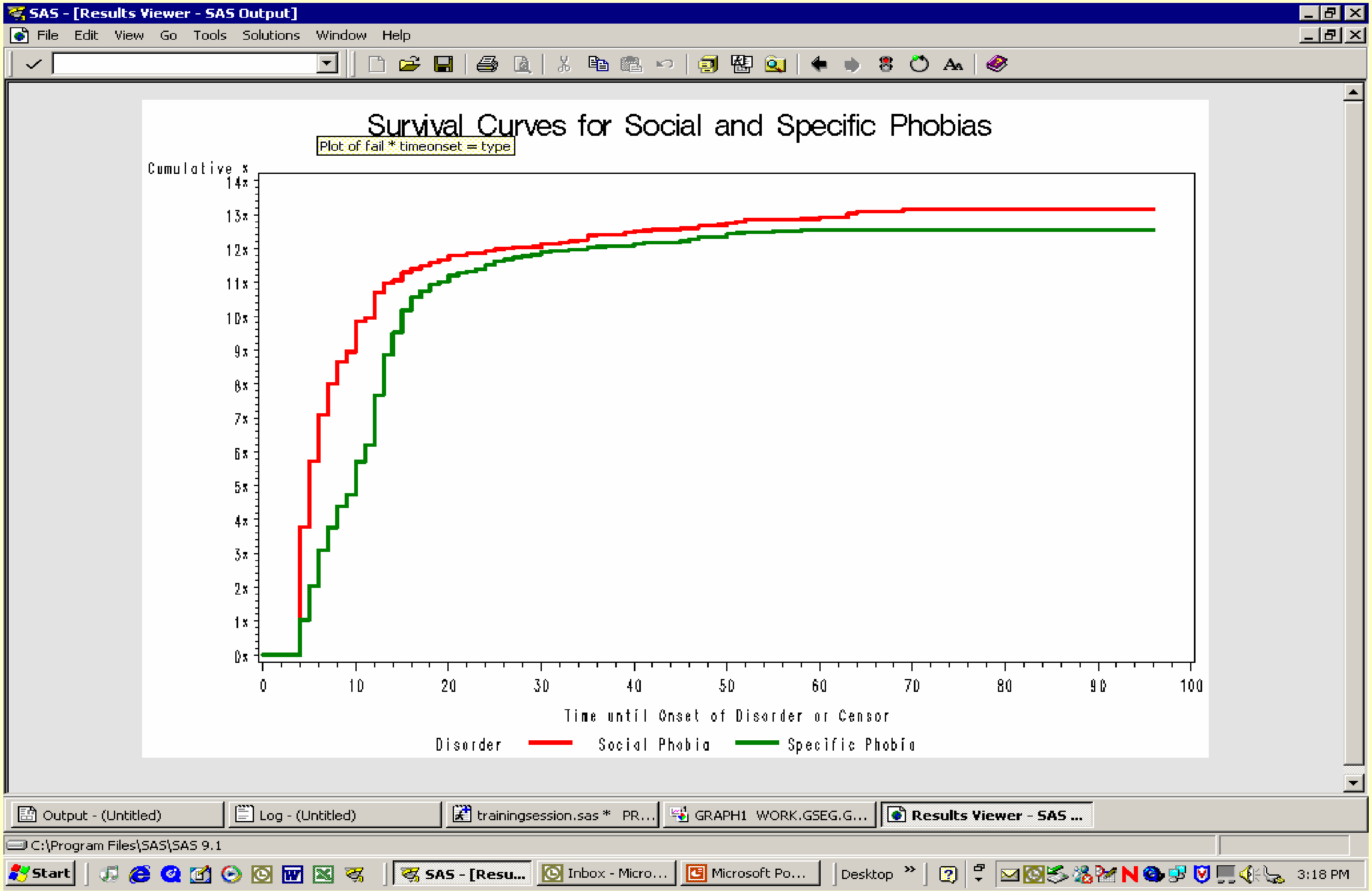
Using SAS ODS to Transfer Analysis Output to External Software

- The Output Delivery System of SAS allows various output delivery destinations such as various files types (HTML, PDF, RTF) as well as output datasets for each part of the procedure output
- Use of ODS can make moving analysis output into software of choice automated and error-free
- SAS graphing and reporting tools are fully capable of all types of reports but many journals request other formats such as Word or PDF files
- ODS offers a number of methods for moving tabular and graphical output into files of a type that Excel can read, HTML, tagsets.msoffice2k, GIF/BMP, and other files types

Example of HTML Output from Output Delivery System

```
ods html style=analysis  ;
proc gplot ;
title "Survival Curves for Social and Specific Phobias"
  ;
      plot fail*timeonset=type  / legend  ;
      format fail percent10. type typef.  ;
      run ;
ods html close ;
```

Example of Using ODS with HTML



Survival Curve Exercises

1. Replicate the examples show today using our code if you are not familiar with SAS coding. Replicate the survival curve with just 1 line first and try doing the 2 lines per graph if you have time.

2. Try creating a curve for a different disorder if you would like to do something new.

Preparing Data for Discrete-Time Logistic Regression

- Create survival dataset using "output" statement in SAS, turns person-level file into person-year file or equivalent multiple record per individual type file
- Create time-varying covariates and dependent variables as well as person-years or "ints"
- Check int*outcome, check all preds*outcome for omitted groups and collapsing
- Check printouts of data to know exactly what is happening with coding, make no assumptions about coding without examination of data

Preparation of the Person-Year Dataset

- Create a person-year or person-day or some type of person-unit of analysis dataset from a person-level dataset, organizes records for correct analysis of timing of event of interest
- This is easily done using a do loop with an output statement in SAS
- For example, we use person-year as our unit of analysis and expand our dataset to a 1 record per person to multiple records per individual, number of records depends on the variables in the do loop

Creating a Multiple Record File from a Single Record File

```
data personyear ;
set two ;
do int = 1 to age ;
output ;
end ;

proc freq ;
title "Distribution of Ints for NCS-R Person Year File" ;
tables int ;
run ;
```

Distribution of Person Years/Ints

Distribution of Ints for NCS-R Person Year File

The FREQ Procedure

int	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	9282	2.24	9282	2.24
2	9282	2.24	18564	4.47
3	9282	2.24	27846	6.71
4	9282	2.24	37128	8.94
5	9282	2.24	46410	11.18
6	9282	2.24	55692	13.41
7	9282	2.24	64974	15.65
8	9282	2.24	74256	17.88
9	9282	2.24	83538	20.12
10	9282	2.24	92820	22.35
11	9282	2.24	102102	24.59
12	9282	2.24	111384	26.83
13	9282	2.24	120666	29.06
14	9282	2.24	129948	31.30
15	9282	2.24	139230	33.53
16	9282	2.24	148512	35.77
17	9282	2.24	157794	38.00
18	9281	2.24	167075	40.24
19	9113	2.19	176188	42.43
20	8927	2.15	185115	44.58
21	8741	2.11	193856	46.69
22	8562	2.06	202418	48.75
23	8383	2.02	210801	50.77
24	8206	1.98	219007	52.75
25	8041	1.94	227048	54.68
26	7860	1.89	234908	56.57
27	7693	1.85	242601	58.43
28	7506	1.81	250107	60.24
29	7344	1.77	257451	62.00
30	7178	1.73	264629	63.73
31	6946	1.67	271575	65.41
32	6782	1.63	278357	67.04
33	6606	1.59	284963	68.63
34	6434	1.55	291397	70.18
35	6248	1.50	297645	71.68
36	6059	1.46	303704	73.14
37	5879	1.42	309583	74.56
38	5678	1.37	315261	75.93
39	5451	1.31	320712	77.24
40	5274	1.27	325986	78.51

41	5047	1.22	331033	79.73
42	4858	1.17	335891	80.90
43	4653	1.12	340544	82.02
44	4443	1.07	344987	83.09
45	4237	1.02	349224	84.11
46	4061	0.98	353285	85.08
47	3917	0.94	357202	86.03
48	3740	0.90	360942	86.93
49	3570	0.86	364512	87.79
50	3383	0.81	367895	88.60
51	3202	0.77	371097	89.37
52	3048	0.73	374145	90.11
53	2867	0.69	377012	90.80
54	2726	0.66	379738	91.46
55	2575	0.62	382313	92.08
56	2460	0.59	384773	92.67
57	2323	0.56	387096	93.23
58	2210	0.53	389306	93.76
59	2074	0.50	391380	94.26
60	1952	0.47	393332	94.73
61	1838	0.44	395170	95.17
62	1745	0.42	396915	95.59
63	1638	0.39	398553	95.99
64	1561	0.38	400114	96.36
65	1461	0.35	401575	96.71
66	1373	0.33	402948	97.04
67	1295	0.31	404243	97.36
68	1213	0.29	405456	97.65
69	1114	0.27	406570	97.92
70	1031	0.25	407601	98.17
71	958	0.23	408559	98.40
72	874	0.21	409433	98.61
73	807	0.19	410240	98.80
74	731	0.18	410971	98.98
75	648	0.16	411619	99.13
76	583	0.14	412202	99.27
77	503	0.12	412705	99.39
78	442	0.11	413147	99.50
79	378	0.09	413525	99.59
80	330	0.08	413855	99.67
81	276	0.07	414131	99.74
82	221	0.05	414352	99.79
83	185	0.04	414537	99.84
84	156	0.04	414693	99.87
85	122	0.03	414815	99.90
86	103	0.02	414918	99.93
87	82	0.02	415000	99.95
88	63	0.02	415063	99.96
89	49	0.01	415112	99.97
90	38	0.01	415150	99.98
91	22	0.01	415172	99.99
92	15	0.00	415187	99.99
93	11	0.00	415198	100.00
94	6	0.00	415204	100.00
95	4	0.00	415208	100.00
96	3	0.00	415211	100.00
97	3	0.00	415214	100.00
98	3	0.00	415217	100.00
99	1	0.00	415218	100.00

Printout of One Person's Person-Level and Person-Year Data Records

From the Person Level File:

Obs	Caseid	Age
74	40100100711	55

From the Person-Year File:

Obs	Caseid	Age	int
1	40100100711	55	1
2	40100100711	55	2
3	40100100711	55	3
4	40100100711	55	4
5	40100100711	55	5
6	40100100711	55	6
7	40100100711	55	7
8	40100100711	55	8
9	40100100711	55	9
10	40100100711	55	10
11	40100100711	55	11
12	40100100711	55	12
13	40100100711	55	13
14	40100100711	55	14
15	40100100711	55	15
16	40100100711	55	16
17	40100100711	55	17
18	40100100711	55	18
19	40100100711	55	19
20	40100100711	55	20
21	40100100711	55	21

Etc to age 55

Create Time-Varying Dependent Variables and Covariates

- Preparation includes creation of time-varying variables as needed for the analysis
- We want a time varying outcome and time varying educational status
- Additionally we need a year of life variable which we call "int" and is created in the output do loop, this variable measures year of life with each person having a value from 1 to the year of interview

Time-Dependent Outcome

- For this analysis we are interested in onset of Major Depressive Disorder and so create the following time dependent outcomes:

- The outcome variables is set to yes or 1 only in the year of onset of substance use with all other person years set to 0

```
**create time varying outcome and education for models* ;
data personyear ;
  set personyear ;
if 4<=mde_ond<=86 and int=mde_ond then mdeonset=1 ; else
  mdeonset=0 ;
proc print data=personyear ;
where mde_ond=4 ;
var caseid int mde_ond mdeonset ;
run ;
```

- This type of coding differs from the person level file since it now not only has a yes for lifetime MDE but also identifies the year in which the onset occurs, this is now the event of interest

Printout of Person Year Records for MDE at Age 4

- This person has 38 records so age at interview was 38 but onset of MDE at age 4, indicated by the 1 in the mdonset variable followed by all zeros for the rest of the data array
- Note that once a person has the event of interest they are no longer at risk and these records will not be used in predicting time to onset of MDE

Obs	CASEID	int	MDE_OND	mdeonset
38023	848	1	4	0
38024	848	2	4	0
38025	848	3	4	0
38026	848	4	4	1
38027	848	5	4	0
38028	848	6	4	0
38029	848	7	4	0
38030	848	8	4	0
38031	848	9	4	0
38032	848	10	4	0
38033	848	11	4	0
38034	848	12	4	0
38035	848	13	4	0
38036	848	14	4	0
38037	848	15	4	0
38038	848	16	4	0
38039	848	17	4	0
38040	848	18	4	0
38041	848	19	4	0
38042	848	20	4	0
38043	848	21	4	0
38044	848	22	4	0
38045	848	23	4	0
38046	848	24	4	0
38047	848	25	4	0
38048	848	26	4	0
38049	848	27	4	0
38050	848	28	4	0
38051	848	29	4	0
38052	848	30	4	0
38053	848	31	4	0
38054	848	32	4	0
38055	848	33	4	0
38056	848	34	4	0
38057	848	35	4	0
38058	848	36	4	0
38059	848	37	4	0
38060	848	38	4	0

Time-Varying Predictor Variables

For the model we will run we use the following predictors to predict MDE Onset:

Time-Invariant predictors:

Sex (sexf)

Age (agecat)

Race (racecat_)

Time-varying predictors:

Education (educattv)

Controls developed as categorical ints or person years:

Ints (continuous but could be categorical)

Creating a Time-Varying Education Covariate

- Educational achievement is currently expressed in terms of years of education completed

- We can convert this to a time varying education variable by adding 6 (usual age of beginning school) to the number of years of education completed and develop a variable as follows:

  1. add 6 to number of years of education completed, <= this number means student
  2. for all person years > yrs of education + 6 are non-student 0 with time invariant years of education

```
*create time - varying education by adding 6 and calling years <= this sum as student* ;
eductv=educ + 6 ;

if eductv >= int then student=1 ; else student=0 ;
if int > eductv and educ_cat=1 then ns0_11=1 ; else ns0_11=0 ;
if int > eductv and educ_cat=2 then ns12=1 ; else ns12=0 ;
if int > eductv and educ_cat=3 then ns13_15=1 ; else ns13_15=0 ;
if int > eductv and educ_cat=4 then ns16=1 ; else ns16=0 ;
```

Check education variable

Obs	CASEID	int	MDE_OND	mdeonset	eductv	EDUC_CAT	EDUC	student	ns0_11	ns12	ns13_15	ns16
38023	848	1	4	0	16	1	10	1	0	0	0	0
38024	848	2	4	0	16	1	10	1	0	0	0	0
38025	848	3	4	0	16	1	10	1	0	0	0	0
38026	848	4	4	1	16	1	10	1	0	0	0	0
38027	848	5	4	0	16	1	10	1	0	0	0	0
38028	848	6	4	0	16	1	10	1	0	0	0	0
38029	848	7	4	0	16	1	10	1	0	0	0	0
38030	848	8	4	0	16	1	10	1	0	0	0	0
38031	848	9	4	0	16	1	10	1	0	0	0	0
38032	848	10	4	0	16	1	10	1	0	0	0	0
38033	848	11	4	0	16	1	10	1	0	0	0	0
38034	848	12	4	0	16	1	10	1	0	0	0	0
38035	848	13	4	0	16	1	10	1	0	0	0	0
38036	848	14	4	0	16	1	10	1	0	0	0	0
38037	848	15	4	0	16	1	10	1	0	0	0	0
38038	848	16	4	0	16	1	10	1	0	0	0	0
38039	848	17	4	0	16	1	10	0	1	0	0	0
38040	848	18	4	0	16	1	10	0	1	0	0	0
38041	848	19	4	0	16	1	10	0	1	0	0	0
38042	848	20	4	0	16	1	10	0	1	0	0	0
38043	848	21	4	0	16	1	10	0	1	0	0	0
38044	848	22	4	0	16	1	10	0	1	0	0	0
38045	848	23	4	0	16	1	10	0	1	0	0	0
38046	848	24	4	0	16	1	10	0	1	0	0	0
38047	848	25	4	0	16	1	10	0	1	0	0	0
38048	848	26	4	0	16	1	10	0	1	0	0	0
38049	848	27	4	0	16	1	10	0	1	0	0	0
38050	848	28	4	0	16	1	10	0	1	0	0	0
38051	848	29	4	0	16	1	10	0	1	0	0	0
38052	848	30	4	0	16	1	10	0	1	0	0	0
38053	848	31	4	0	16	1	10	0	1	0	0	0
38054	848	32	4	0	16	1	10	0	1	0	0	0
38055	848	33	4	0	16	1	10	0	1	0	0	0
38056	848	34	4	0	16	1	10	0	1	0	0	0
38057	848	35	4	0	16	1	10	0	1	0	0	0
38058	848	36	4	0	16	1	10	0	1	0	0	0
38059	848	37	4	0	16	1	10	0	1	0	0	0
38060	848	38	4	0	16	1	10	0	1	0	0	0

Variable Checking and Collapsing

- Often you will need to consider a collapsing scheme or omitting strategy with the person years variable (int) or other predictors
- Systematic examination of cross tabs of the outcome and the predictors, within the sample you will model in is essential to getting your models to converge properly
- Indications of problems with convergence are huge OR's or standard errors even though model apparently converges, no convergence and other odd results that are meaningless

Defining Years at Risk

- Ints can be used as either single year dummy variables or as collapsed variables
- Similar checking with outcome*(all other predictors) should be done as well so you can use a well-educated model formulation strategy
- Other issues are that you will want to consider the years of risk and what particular years of risk should be included in the models
- In the case of modeling the outcome of MDE, you want the years of year 1 of life to year of event occurring (age of onset of MDE) or year censored (age of interview), note that once the outcome occurs that person is no longer at "risk" and person years after that point are no longer included in the analysis

Discrete-Time Logistic Regression

- Here is a simple example of a discrete time logistic regression using PROC SURVEYLOGISTIC
- Note the where statement selecting person years from age 1 to either event (onset of MDE)/age at censor (age at interview)
- Also note that the "ints" or years of life are included as a categorical predictor, need these to represent time units in model
- This code first creates the categorical ints

```
**collapsing ints into categories* ;
if 1<=int<=12 then intcat=1 ;
else if 13<=int<=19 then intcat=2 ;
else if 20<=int<=29 then intcat=3 ;
else if 30<=int<=39 then intcat=4 ;
else intcat=5 ;
```

SURVEYLOGISTIC Code

```
proc surveylogistic ;
options ls=119 ps=60 ;
strata str ;
cluster secu ;
weight finalp1w ;
class agecat sex mar_stat region racecat_ intcat  /
  param=reference ;
model mdeonset (event='1') = intcat agecat sex mar_stat
  region racecat_  student ns0_11 ns12 ns13_15  ;
where int <= ageevent  ;
format agecat agef. intcat intf. ;
run ;
```

SURVEYLOGISTIC Results

The SURVEYLOGISTIC Procedure

Model Information

Data Set	WORK.PERSONYEAR	
Response Variable	mdeonset	
Number of Response Levels	2	
Stratum Variable	STR	Strata NCS-R version
Number of Strata	42	
Cluster Variable	SECU	Sampling error computation unit
Number of Clusters	84	
Weight Variable	FINALP1W	Final part 1 weight
Model	Binary Logit	
Optimization Technique	Fisher's Scoring	
Variance Adjustment	Degrees of Freedom (DF)	

Number of Observations Read	385696
Number of Observations Used	385696
Sum of Weights Read	386871.6
Sum of Weights Used	386871.6

Response Profile

Ordered Value	mdeonset	Total Frequency	Total Weight
1	0	383867	385092.28
2	1	1829	1779.36

Probability modeled is mdeonset=1.

Results, continued

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	22704.982	21519.381
SC	22715.845	21747.500
-2 Log L	22702.982	21477.381

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1225.6015	20	<.0001
Score	1139.0336	20	<.0001
Wald	2180.6682	20	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
intcat	4	244.5714	<.0001
agecat	3	361.4684	<.0001
SEX	1	53.4344	<.0001
MAR_STAT	2	54.8093	<.0001
REGION	3	13.5039	0.0037
RACECAT_	3	39.2688	<.0001
student	1	0.5952	0.4404
ns0_11	1	0.2931	0.5882
ns12	1	4.4792	0.0343
ns13_15	1	12.7798	0.0004

Results, continued

The SURVEYLOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.0351	0.1098	1349.9545	<.0001
intcat	1-12	1	-1.2427	0.2170	32.7860	<.0001
intcat	13-19	1	0.0242	0.2172	0.0124	0.9113
intcat	20-29	1	-0.3458	0.1025	11.3821	0.0007
intcat	30-39	1	-0.0864	0.0794	1.1854	0.2763
agecat	30-44	1	-0.5066	0.0885	32.7612	<.0001
agecat	45-59	1	-1.0134	0.1009	100.9462	<.0001
agecat	60+	1	-2.3309	0.1368	290.3981	<.0001
SEX	(0) FEMALE	1	0.4626	0.0633	53.4344	<.0001
MAR_STAT	(1) MARRIED/COHABITATING	1	-0.2528	0.0964	6.8777	0.0087
MAR_STAT	(2) SEPARATED/WIDOWED/DIVORCED	1	0.2128	0.1026	4.3014	0.0381
REGION	(1) NORTHEAST	1	-0.1621	0.0872	3.4588	0.0629
REGION	(2) MIDWEST	1	-0.1212	0.0812	2.2266	0.1357
REGION	(3) SOUTH	1	-0.2232	0.0614	13.2081	0.0003
RACECAT_	(1) HISPANIC	1	-0.3415	0.0931	13.4680	0.0002
RACECAT_	(2) BLACK	1	-0.5423	0.0982	30.5011	<.0001
RACECAT_	(3) OTHER	1	-0.1306	0.1186	1.2126	0.2708
student		1	-0.1468	0.1903	0.5952	0.4404
ns0_11		1	0.0586	0.1082	0.2931	0.5882
ns12		1	0.1504	0.0711	4.4792	0.0343
ns13_15		1	0.3188	0.0892	12.7798	0.0004

Results, continued

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
intcat 1-12 vs 40+	0.289	0.189	0.442
intcat 13-19 vs 40+	1.024	0.669	1.568
intcat 20-29 vs 40+	0.708	0.579	0.865
intcat 30-39 vs 40+	0.917	0.785	1.072
agecat 30-44 vs <=29	0.603	0.507	0.717
agecat 45-59 vs <=29	0.363	0.298	0.442
agecat 60+ vs <=29	0.097	0.074	0.127
SEX (0) FEMALE vs (1) MALE	1.588	1.403	1.798
MAR_STAT (1) MARRIED/COHABITATING vs (3) NEVER MARRIED	0.777	0.643	0.938
MAR_STAT (2) SEPARATED/WIDOWED/DIVORCED vs (3) NEVER MARRIED	1.237	1.012	1.513
REGION (1) NORTHEAST vs (4) WEST	0.850	0.717	1.009
REGION (2) MIDWEST vs (4) WEST	0.886	0.756	1.039
REGION (3) SOUTH vs (4) WEST	0.800	0.709	0.902
RACECAT_ (1) HISPANIC vs (4) WHITE	0.711	0.592	0.853
RACECAT_ (2) BLACK vs (4) WHITE	0.581	0.480	0.705
RACECAT_ (3) OTHER vs (4) WHITE	0.878	0.696	1.107
student	0.863	0.595	1.254
ns0_11	1.060	0.858	1.311
ns12	1.162	1.011	1.336
ns13_15	1.376	1.155	1.638

- Young people <= 29 years at interview are more likely than older cohorts to have MDE
- Women are more likely than men to have MDE
- Being separated/widowed or divorced significantly predicts MDE as compared to never having married
- Respondents living in regions other than the West are signficantly less likely to have onset of MDE
- Hispanics and Black are less likely than Whites to have MDE
- Respondents with 12-15 years of education (non-students) are more likely to have MDE compared to non-students with 16+ years of education

Summary of Discrete-Time Logistic Regression

- We have presented just 1 example of how to set up a dataset for a person-year format, create time-varying variables as both outcome and predictors, and specify the model correctly

- These concepts can obviously be extended to include interactions, linear contrasts and sub-population analyses

- Time does not permit extensive examples in this area but many of the Kessler et al publications includes analysis techniques of this type

Discrete-Time Logistic Exercises

1. Either replicate the model presented in this section by first creating the person year dataset, then creating the variables needed, and then running the model as specified.

2. For those of you with SAS experience feel free to try other predictors in the model or try developing another time varying outcome such as onset of Social phobia instead of MDE.

General Question and Answer Session

- During the last part of the training, we will gather questions of both general and individual nature and try to address them either within the group or individually.
- We will develop a list of questions and then decide how to best address your concerns.

Resources and References

- NCS-R website: FAQ section: we have developed a general FAQ section with posted answers and will continue to post QA as they come in
- SAMDHA website for technical/documentation questions

- SAS website: http://sas.com/
- Sudaan website:http://www.rti.org/sudaan/
- SAS documentation located at http://sas.com/ under products and solutions, documentation
- Sudaan documentation located at http://www.rti.org/sudaan/ under documentation area

References

References for data preparation and descriptive analysis

Cody, R.P. and Smith, J.K., "Applied Statistics and the SAS Programming Language", Cary, NC: SAS Institute Inc.

Delwiche, L. and Slaughter, S, "The Little SAS Book, a Primer", Cary, NC: SAS Institute Inc.

The SAS Language Guide, Reference, SAS Institute Publishing

References for Analysis of Categorical data:

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Stokes, M.E., Davis, C.S., and Koch, G.G. (2000), *Categorical Data Analysis Using the SAS System, Second Edition*, Cary, NC: SAS Institute Inc.

References for Survival Analysis

Allison, P.D., Survival Analysis using the SAS System, A Practical Guide, Cary, NC: SAS Institute Inc

Allison, P.D. 1982 "Discrete-time methods for the analysis of event histories." Pp. 61-98 in Samuel Leinhardt (ed.), *Sociological Methodology 1982*. San Francisco: Jossey-Bass

Bradley Efron, Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve Journal**,** J Amer Stat Assn, Volume 83, pages 414-425 1988

References for analysis of complex survey data

Heeringa, S., and Liu, J. (1997), Complex sample design effects and inference for mental health survey data, *International Journal of Methods in Psychiatric Research, 7*, Whurr Publishers Ltd. – Pages 221 – 230.

Rust, K., (1985), Variance Estimation for Complex Estimators in Sample Surveys, *Journal of Official Statistics, 1 (4)*, Statistics Sweden Publishing Service. Pages 381 –397

Landis, J., Lepkowski, J., Eklund, S., and Stehouwer, S., (1984) A Statistical Methodology for Analyzing Data from a Complex Survey: The first National Health and Nutrition Examination Survey, *Vital Health Statistics, 21 (10)*, MD: U.S. National Center for Health Statistics.

Little, R.J.A., Lewitzky, S., Heeringa, S., Lepkowski, J., and Kessler, R.C., (1997), Assessment of Weighting Methodology for the National Comorbidity Survey, *American Journal of Epidemiology, 146 (5)*. –Pages 439 – 449.

Reference for Linear Regression:

Freund, R.J. and Littell, R.C. (1986), *SAS System for Regression, 1986 Edition*, Cary, NC: SAS Institute Inc