MULTILOG Example #1

SUDAAN Statements and Results Illustrated

- Logistic regression modeling
- R and SEMETHOD options
- CONDMARG
- ADJRR option
- CATLEVEL

Input Data Set(s): DARE.SSD

Example

Evaluate the effect of exposure to a Drug Abuse Prevention Program (Project DARE) by fitting logistic regression models in MULTILOG. The dependent variable in the analysis is cigarette initiation, and the population of interest is students in the 5th and 6th grade. In addition to exposure to the DARE program, also consider grade in school, gender, race/ethnicity, family composition, and metropolitan status as independent variables. Compare fitted models assuming independent and exchangeable working correlations.

This example highlights the use of the PROC statement options (R and SEMETHOD) to implement GEE model-fitting techniques for cluster-correlated experimental data.

This example also highlights the estimation of the model-adjusted risk and risk ratio for smoking initiation via conditional marginal proportions (ADJRR option on CONDMARG statement). Confidence intervals for the model-adjusted risk are new in Release 11.0.

Solution

Experimental studies of the effect of prevention programs on substance use are often based on nested cohort designs in which intact social groups or clusters of individuals are randomized to treatment conditions, and individuals within the clusters are followed over time as a cohort to evaluate the effects of treatment. The units of assignment may be schools, communities, or worksites, but the units of observation are the students, community residents, or workers. Because they are exposed to a common set of circumstances, students within the same school tend to be positively correlated with one another. This positive intracluster correlation implies that the observational units are no longer statistically independent. Unless the intracluster correlation that results from the sampling design is accounted for in the statistical analysis, estimated standard errors of the treatment effects will generally be underestimated, leading to inflated Type I error rates and false-positive tests of treatment effects (Murray and Hannan, 1990; Moskowitz, Malvin, Schaeffer, and Schaps, 1984; Donner, 1982; Donner, Birkett, and Buck, 1981).

Illustrative data for this example were collected as part of a longitudinal evaluation of Project DARE (Drug Abuse Resistance Education) on substance abuse outcomes in Illinois (Ennett, Rosenbaum, Flewelling, Bieler, Ringwalt, and Bailey, 1994).

Exhibit 1 shows the structure of the data used in this example.

Exhibit 1. Structure of the DARE Data

Exposure Group 1 = Control, 2 = DARE	School ID (Cluster)	Student ID	Y = cigarette initiation 1= yes, 2 = no
1	1	1	2
1	1	2	1
1	1	3	2
1	2	1	2
1	2	2	2
2	10	1	2
2	10	2	1
2	20	1	1
2	20	2	1
2	30	1	1
.	.	.	.
.	.	.	.
.	.	.	.

N = 1,525 records on the file (1,525 students clustered within 36 schools)

In this example, we analyze a single dependent variable that is representative of outcome measures used to evaluate drug use prevention programs. At each wave of data collection, students were asked whether they had ever smoked cigarettes. The binary dependent variable relates to the initiation of cigarette use between Waves 1 and 2 (coded 1 if the adolescent initiated cigarette use; 2 = otherwise). The desired effect is a negative correlation with DARE (coded 1 = adolescent exposed to DARE; 2 = not exposed). The sample for initiation analysis is limited to students who reported no lifetime use at Wave 1.

We report results for the covariate of primary interest, exposure to the DARE program, as well as the following background characteristics (with 8 degrees of freedom): grade in school, sex, race/ethnicity, family composition, and metropolitan status. Respondents included 34% fifth-grade and 66% sixth-grade students; approximately half were male. The sample was 51% white, 24% African-American, 9% Hispanic, and 16% "other." The majority (65%) lived with both parents in the same household. Fewer respondents lived in rural areas (26%) compared with suburban (38%) and urban (36%) areas.

We used SUDAAN's MULTILOG procedure to fit a logistic regression model to the binary response variable of interest via the GEE model-fitting method, under both independent and exchangeable working correlations. The independence working assumption amounts to ordinary logistic regression. The use of the variance correction (standard in SUDAAN) yields valid results in the presence of intracluster correlation. In fact, the robust variance estimate ensures that the results are robust to any misspecification of the correlation structure. We also provide results using the model-based variance estimates. In ***Exhibit 2***, we compare four different ways of fitting the model, all of which can be implemented in MULTILOG: independent working correlations, with and without a variance correction; and exchangeable working correlations, with and without a variance correction.

Using SUDAAN, the DARE program is shown to have a significant negative effect on the initiation of cigarette use, regardless of the working assumptions about the correlation structure (p=0.0369 under working independence; p=0.0216 under exchangeability). The estimated intracluster correlation under exchangeability is 0.0206. Use of a robust variance estimate ensures that the results of statistical analyses are valid regardless of what the true correlation structure is. In this example, the exchangeability assumption appears to be correct, since results using the robust and model-based variance estimates were essentially the same. The advantage of modeling the correlation structure (e.g., through exchangeability) is its potential to improve efficiency, and thereby increasing the power of statistical analyses.

Exhibit 2. Evaluation of the DARE Effect on Cigarette Initiation Via Logistic Regression Modeling in MULTILOG

Variable	Statistic	Working Correlations			
		Independent (Ordinary Logistic Regression)		Exchangeable	
		No Variance Correction	Variance Correction	No Variance Correction	Variance Correction
Initiation of Cigarette Use By Wave 2	β	-0.5225	-0.5225	-0.5825	-0.5825
	SE	0.1821	0.2408	0.2433	0.2422
	Observed DEFF	--	1.75	--	--
	Z-statistic	-2.87	-2.17	-2.39	-2.41
	P-value	**0.0069**	**0.0369**	**0.0221**	**0.0216**
		Model-Based	Zeger or Binder	Model-Based	Zeger or Binder

Working Correlations **SUDAAN PROC MULTILOG Settings**

Independent (Ordinary Logistic Regression)
No variance correction R=independent, SEMETHOD=Model
Variance correction (robust variance) R=independent, SEMETHOD=Zeger

Exchangeable
No variance correction (model-based [naive] variance) R=exchangeable, SEMETHOD=Model
Variance Correction (robust variance) R=exchangeable, SEMETHOD=Zeger

The unadjusted incidence of cigarette use during the intervention was significantly lower among students who participated in DARE (9.5% observed for DARE vs. 15.4% for controls). As seen in ***Exhibit 2***, naively ignoring the intracluster correlation leads to a much more significant treatment effect (p=0.0069). The observed design effect for DARE was 1.75, which indicates almost a doubling in the variance of the estimated treatment effect under cluster randomization.

The following sets of programming statements fit different versions of a logistic model in SUDAAN PROC MULTILOG. Since there is no DESIGN option specified on the PROC statement, SUDAAN is using the default DESIGN=WR (with replacement) option for variance estimation.

To obtain the results in ***Exhibit 2***, we fit the following four types of GEE logistic regression models in MULTILOG:

- **R=INDEPENDENT and SEMETHOD=ZEGER**—Implements the GEE model-fitting technique under an independent "working" assumption and Zeger and Liang's (1986) robust variance estimator. This model is sometimes referred to as ordinary logistic regression with a variance correction. Note that for binary outcomes, SEMETHOD=ZEGER is equivalent to SEMETHOD=BINDER.
- **R=INDEPENDENT and SEMETHOD=MODEL**—This is ordinary logistic regression *without* a variance correction. Literally, this combination implies an independent "working" assumption and a model-based or naive variance estimator. The variance estimator is naive in the sense that it computes variances as if the independence working assumption were correct. *This option is <u>not</u> valid for cluster-correlated data and is presented only for comparison purposes.*

- **R=EXCHANGEABLE and SEMETHOD=ZEGER**—Implements the GEE model-fitting technique under exchangeable "working" correlations and Zeger and Liang's (1986) robust variance estimator.
- **R=EXCHANGEABLE and SEMETHOD=MODEL**—Implements the GEE model-fitting technique under exchangeable "working" correlations and a model-based variance estimator. Variances are computed as if the exchangeable "working" correlation assumption were correct.

In this example, the NEST statement indicates that SCHOOL is the cluster variable. The WEIGHT statement indicates equal sampling weights of 1.0 for each student on the file.

In MULTILOG, the CLASS statement contains the dependent variable *and* all covariates that are to be modeled as categorical covariates.

The MODEL statement specifies the categorical dependent variable INTCIG12 on the left of the "=" sign (with levels 1 and 2), and independent regressors on the right:

- DARE (1=Exposed to DARE, 2=Not Exposed);
- FIFTH (1=5th Grade, 2=6th Grade);
- SEX (1=Males, 2=Females);
- RACE (1=Black, 2=Hispanic, 3=Other, 4=White);
- OTHFAM (1=Non-Traditional; 2=Traditional); and
- AREA (1=Rural, 2=Suburban, 3=Urban).

For binary responses, the CUMLOGIT (cumulative logit) and GENLOGIT (generalized logit) links specify the same logistic regression model. The default Wald-*F* test is used for all tests of hypotheses.

The CONDMARG statement requests the conditional marginal proportion (*model-adjusted risk*) for each level of DARE exposure. The log odds of cigarette initiation for a given level of DARE exposure are calculated from the estimated linear model by specifying the value of the DARE variable as the level of interest and then by specifying all other variables in the model (except DARE) to be the estimated percentage distribution in the population. Based on the obtained log odds, the probability of cigarette initiation (model-adjusted risk) is then calculated for a specific level of the DARE variable. The ADJRR option on the CONDMARG statement computes the *model-adjusted risk ratio* for DARE level 1 vs. 2 (exposed vs. not exposed).

We include multiple PRINT statements, all of which are optional. Multiple PRINT statements allow us to set up different default print environments (SETENV statements) for different PRINT groups. The PRINT statements are used in this example to request the PRINT groups of interest; individual statistics of interest, and in some cases, to change default labels for those statistics; and to specify a variety of formats for those printed statistics. Without the PRINT statement, default statistics are produced from each PRINT group, with default formats.

The SETENV statements are optional. They set up default formats for printed statistics and further manipulate the printout to the needs of the user.

The RFORMAT statements associate the SAS formats with the variables used in the procedure. The RLABEL statement defines variable labels for use in the current procedure only. Without the RLABEL statement, SAS variable labels would be produced if already defined.

This example begins with the DESCRIPT procedure to estimate the incidence of cigarette smoking in each DARE group (exposed vs. unexposed). These percentages are unadjusted for any other covariates. Just as in MULTILOG, the NEST statement indicates that SCHOOL is the cluster variable. The WEIGHT statement indicates equal sampling weights of 1.0 for each student on the file.

The response variable INTCIG12 appears on the VAR statement. The CATLEVEL statement specifies that we want to estimate percentages for INTCIG12=1. The TABLES statement requests the estimated percentage for each level of DARE—exposed vs. unexposed.

This example was run in SAS-Callable SUDAAN, and the SAS program and *.LST files are provided.

Exhibit 3. SAS-Callable SUDAAN Code (PROC DESCRIPT)

```
libname in v604 "c:\10winbetatest\examplemanual\multilog";

options nocenter pagesize=70 linesize=85;
proc format;
  value dare 1="1=Exposed"
             2="2=Not Exposed";
  value yesno 1="1=Yes"
              2="2=No";
  value grade 1="1=5th Grade"
              2="2=6th Grade";
  value sex 1="1=Male"
           2="2=Female";
  value race 1="1=Black"
             2="2=Hispanic"
             3="3=Other"
             4="4=White";
  value family 1="1=Non-Traditional"
               2="2=Traditional";
  value area 1="1=Rural"
             2="2=Suburban"
             3="3=Urban";

data one; set in.dare;
proc sort data=one; by school;

PROC DESCRIPT DATA=one FILETYPE=SAS NOMARG;
  NEST _ONE_ SCHOOL;
  WEIGHT _ONE_;

  CLASS DARE;
  TABLES DARE;
  VAR INTCIG12;
  CATLEVEL 1;

  SETENV LABWIDTH=35 COLWIDTH=6 DECWIDTH=2;
  PRINT NSUM PERCENT SEPERCENT="SE" DEFFPCT="Design Effect" /
        NSUMFMT=F6.0 PERCENTFMT=F7.2 STYLE=NCHS;
  RFORMAT DARE dare.;
  RFORMAT INTCIG12 yesno.;
  RLABEL DARE="DARE Program";
  RLABEL INTCIG12="Initiation of Cigarette Use";
  RTITLE "Descriptive Statistics for Initiation of Cigarette Use";
```

Exhibit 4. First Page of PROC DESCRIPT Output

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute    February 2011
                                Release 11.0.0


DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method,
Assuming a With Replacement (WR) Design
    Sample Weight: _ONE_
    Stratification Variables(s): _ONE_
    Primary Sampling Unit: SCHOOL


Number of observations read    :   1525    Weighted count :     1525
Denominator degrees of freedom :     35
```

Exhibit 5. CLASS Variable Frequencies (DARE)

```
Frequencies and Values for CLASS Variables

by: DARE Program.
-----------------------------------------
DARE Program    Frequency            Value
-----------------------------------------
Ordered
  Position:
  1                   822        1=Exposed
Ordered
  Position:
  2                   703    2=Not Exposed
-----------------------------------------
```

Exhibit 6. Estimated Percentage Distribution (PROC DESCRIPT)

```
Variance Estimation Method: Taylor Series (WR)

Descriptive Statistics for Initiation of Cigarette Use

by: Variable, DARE Program.
-------------------------------------------------------------------------
Variable
   DARE Program                        Sample                      Design
                                       Size     Percent       SE   Effect
-------------------------------------------------------------------------
Initiation of Cigarette Use: 1=Yes
   1=Exposed                            649      9.55     1.77     2.35
   2=Not Exposed                        539     15.40     2.25     2.10
-------------------------------------------------------------------------
```

A total of 1,525 observations were read from the dataset, but only 1,188 are included in the Initiation of Cigarette Use table above, due to missing values for this variable. These results, unadjusted for other model covariates, indicate that 15.4% of students not participating in DARE initiated cigarette smoking during the time of the intervention (***Exhibit 6***), compared to 9.5% of those exposed to DARE. The standard errors estimated by SUDAAN use a between-cluster variance formula and are, therefore, adjusted for clustering. The design effects indicate that the variances of the percentages are more than doubled under cluster randomization.

Next, we use the MULTILOG procedure to find out whether the observed difference is statistically significant after adjustment for other covariates (***Exhibit 7***).

We present output from two MULTILOG logistic models that are reasonable analytic methods for this type of data: 1) R=Independent, SEMETHOD= Zeger, and 2) R=Exchangeable, SEMETHOD=Model.

This example was run in SAS-Callable SUDAAN, and the SAS program and *.LST files are provided.

Exhibit 7. MULTILOG Code (R=Independent, SEMETHOD=Zeger)

```
PROC MULTILOG DATA=one FILETYPE=SAS SEMETHOD=ZEGER R=INDEPENDENT;
  NEST _ONE_ SCHOOL;
  WEIGHT _ONE_;

  CLASS DARE FIFTH SEX RACE OTHFAM AREA INTCIG12;
  MODEL INTCIG12 = DARE FIFTH SEX RACE OTHFAM AREA / CUMLOGIT;
  CONDMARG DARE / adjrr;

  SETENV LABWIDTH=28 COLWIDTH=7 DECWIDTH=4 COLSPCE=2 TOPMGN=0;
  PRINT beta sebeta deft="Design Effect" t_beta p_beta /
        risk=default tests=default
        deftfmt=f6.2 orfmt=f5.3 loworfmt=f9.3 uporfmt=f9.3
        t_betafmt=f6.2 waldffmt=f6.2 dffmt=f7.0;

  SETENV LABWIDTH=22 COLWIDTH=6 DECWIDTH=4 COLSPCE=1 TOPMGN=0;
  PRINT / COND_MRG=default condmrgfmt=f11.4 t_cndmrgfmt=f8.2 p_cndmrgfmt=f8.4;

  SETENV LABWIDTH=40 DECWIDTH=3 COLSPCE=1 TOPMGN=0;
  PRINT / CONDRISK=default cond_rrfmt=f8.3;

  RLABEL DARE="DARE Program";
  RLABEL INTCIG12="Initiation of Cigarette Use";
  RLABEL FIFTH="Grade in School";
  RLABEL OTHFAM="Family Situation";
  RFORMAT DARE dare.;
  RFORMAT INTCIG12 yesno.;
  RFORMAT FIFTH grade.;
  RFORMAT SEX sex.;
  RFORMAT RACE race.;
  RFORMAT OTHFAM family.;
  RFORMAT AREA area.;
  RTITLE "MULTILOG Logistic Regression Model for the DARE Evaluation Study"
         "Ennett, et al, 1994";
```

Exhibit 8. First Page of SUDAAN Output (R=Independent, SEMETHOD=Zeger)

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute      August 2011
                               Release 11.0.0


DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method,
Assuming a With Replacement (WR) Design
    Sample Weight: _ONE_
    Stratification Variables(s): _ONE_
    Primary Sampling Unit: SCHOOL


Independence parameters have converged in 5 iterations

Number of observations read       :   1525    Weighted count:     1525
Observations used in the analysis :   1188    Weighted count:     1188
Denominator degrees of freedom    :     35


Maximum number of estimable parameters for the model is 10

File ONE contains   36 Clusters
  36 clusters were used to fit the model
Maximum cluster size is 123 records
Minimum cluster size is   8 records


Sample and Population Counts for Response Variable INTCIG12
Based on observations used in the analysis
  1=Yes:  Sample Count      145    Population Count       145
  2=No :  Sample Count     1043    Population Count      1043
```

Exhibit 8 indicates that there are 1,525 students (one record per student) on the file, and 1,188 were used in the analysis (337 students were deleted due to missing values on one or more MODEL statement variables). There are 36 clusters (schools), with cluster sizes ranging from 8 to 123. Overall, 145 students reported having initiated cigarette use during the intervention, while 1,043 did not.

Below (***Exhibit 9*** to ***Exhibit 15***) are the individual frequency distributions for the CLASS variables. The only missing values in the model are from the dependent variable, Initiation of Cigarette Use. The distributions for all other variables sum to 1,525.

Exhibit 9. CLASS Variable Frequencies: DARE

```
Frequencies and Values for CLASS Variables
by: DARE Program.

-----------------------------------------
DARE Program    Frequency            Value
-----------------------------------------
Ordered
  Position:
  1                   822        1=Exposed
Ordered
  Position:
  2                   703    2=Not Exposed
-----------------------------------------
```

Exhibit 10. CLASS Variable Frequencies: Grade in School

```
Frequencies and Values for CLASS Variables

by: Grade in School.
----------------------------------------
Grade in
  School        Frequency          Value
----------------------------------------
Ordered
  Position:
  1                   526    1=5th Grade
Ordered
  Position:
  2                   999    2=6th Grade
----------------------------------------
```

Exhibit 11. CLASS Variable Frequencies: Sex

```
Frequencies and Values for CLASS Variables

by: SEX.
-------------------------------------
SEX             Frequency       Value
-------------------------------------
Ordered
  Position:
  1                   779      1=Male
Ordered
  Position:
  2                   746    2=Female
-------------------------------------
```

Exhibit 12. CLASS Variable Frequencies: Race

```
Frequencies and Values for CLASS Variables

by: RACE.
--------------------------------------
RACE            Frequency         Value
--------------------------------------
Ordered
  Position:
  1                   362       1=Black
Ordered
  Position:
  2                   139    2=Hispanic
Ordered
  Position:
  3                   242       3=Other
Ordered
  Position:
  4                   782       4=White
--------------------------------------
```

Exhibit 13. CLASS Variable Frequencies: Family Situation

```
Frequencies and Values for CLASS Variables

by: Family Situation.
-----------------------------------------------
Family
  Situation     Frequency                Value
-----------------------------------------------
Ordered
  Position:
  1                   533    1=Non-Traditional
Ordered
  Position:
  2                   992        2=Traditional
-----------------------------------------------
```

Exhibit 14. CLASS Variable Frequencies: Area

```
Frequencies and Values for CLASS Variables

by: AREA.
---------------------------------------
AREA            Frequency         Value
---------------------------------------
Ordered
  Position:
  1                   393       1=Rural
Ordered
  Position:
  2                   574    2=Suburban
Ordered
  Position:
  3                   558       3=Urban
---------------------------------------
```

Exhibit 15. CLASS Variable Frequencies: Initiation of Cigarette Use

```
Frequencies and Values for CLASS Variables

by: Initiation of Cigarette Use.
---------------------------------
Initiation
  of
  Cigarette
  Use           Frequency    Value
---------------------------------
Ordered
  Position:
  1                   145    1=Yes
Ordered
  Position:
  2                  1043     2=No
---------------------------------
```

Exhibit 16. Estimated Regression Coefficients (R=Independent, SEMETHOD=Zeger)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Zeger-Liang, 1986)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

INTCIG12 (cum-logit), Independent Variables and Effects	Beta Coeff.	SE Beta	DESIGN EFFECT	T-Test B=0	P-value T-Test B=0
INTCIG12 (cum-logit)					
Intercept 1: 1=Yes	-1.8476	0.4659	1.99	-3.97	0.0003
DARE Program					
1=Exposed	-0.5225	0.2408	1.75	-2.17	0.0369
2=Not Exposed	0.0000	0.0000	.	.	.
Grade in School					
1=5th Grade	-0.5002	0.2494	1.25	-2.01	0.0527
2=6th Grade	0.0000	0.0000	.	.	.
SEX					
1=Male	0.0840	0.1599	0.78	0.53	0.6027
2=Female	0.0000	0.0000	.	.	.
RACE					
1=Black	0.4971	0.3786	1.78	1.31	0.1977
2=Hispanic	0.0951	0.4670	1.46	0.20	0.8398
3=Other	0.4936	0.4214	2.23	1.17	0.2494
4=White	0.0000	0.0000	.	.	.
Family Situation					
1=Non-Traditional	0.4208	0.1706	0.78	2.47	0.0187
2=Traditional	0.0000	0.0000	.	.	.
AREA					
1=Rural	-0.0788	0.3962	1.53	-0.20	0.8435
2=Suburban	-0.2508	0.3610	1.77	-0.69	0.4918
3=Urban	0.0000	0.0000	.	.	.

Exhibit 16 presents the estimated regression coefficient vector, the estimated robust standard errors, design effects, *t*-statistics, and *p*-values for testing H_0: $\beta=0$. Using the GEE-independent approach after adjusting for other covariates in the model, the treatment effect (DARE) is observed to significantly reduce the incidence of cigarette initiation ($p=0.0369$). Other than the treatment effect, only family situation is a statistically significant covariate ($p=0.0187$). The observed design effect for the treatment parameter is 1.75, indicating a 75% increase in variance due to cluster randomization.

Exhibit 17. ANOVA Table (R=Independent, SEMETHOD=Zeger)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Zeger-Liang, 1986)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	10	31.32	0.0000
MODEL MINUS INTERCEPT	9	3.14	0.0071
DARE	1	4.71	0.0369
FIFTH	1	4.02	0.0527
SEX	1	0.28	0.6027
RACE	3	0.63	0.5981
OTHFAM	1	6.08	0.0187
AREA	2	0.30	0.7439

Exhibit 17 presents the statistical significance of all model terms. The default Wald-*F* test is used to evaluate these effects. The *p*-value corresponding to the DARE effect is identical to the regression coefficient table, since this is a 1 df test.

Exhibit 18. Default Odds Ratios (R=Independent, SEMETHOD=Zeger)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Zeger-Liang, 1986)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

INTCIG12 (cum-logit), Independent Variables and Effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
INTCIG12 (cum-logit)			
Intercept 1: 1=Yes	0.158	0.061	0.406
DARE Program			
1=Exposed	0.593	0.364	0.967
2=Not Exposed	1.000	1.000	1.000
Grade in School			
1=5th Grade	0.606	0.366	1.006
2=6th Grade	1.000	1.000	1.000
SEX			
1=Male	1.088	0.786	1.505
2=Female	1.000	1.000	1.000
RACE			
1=Black	1.644	0.762	3.546
2=Hispanic	1.100	0.426	2.838
3=Other	1.638	0.696	3.854
4=White	1.000	1.000	1.000
Family Situation			
1=Non-Traditional	1.523	1.077	2.154
2=Traditional	1.000	1.000	1.000
AREA			
1=Rural	0.924	0.414	2.066
2=Suburban	0.778	0.374	1.619
3=Urban	1.000	1.000	1.000

Exhibit 18 presents the estimated odds ratios and their 95% confidence limits for each regression coefficient in the model. We see that the negative regression coefficient for DARE corresponds to an odds ratio for smoking initiation of 0.593, indicating a protective effect of the DARE program (the odds are reduced by around 40% in the DARE group). Again, each regression coefficient is adjusted for all others in the model.

Exhibit 19. Conditional Marginal Proportions (R=Independent, SEMETHOD=Zeger)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Zeger-Liang, 1986)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

Initiation of Cigarette Use: Conditional Marginal #1	Conditional Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
1=Yes						
DARE Program						
1=Exposed	0.0911	0.0161	0.0632	0.1296	5.64	0.0000
2=Not Exposed	0.1445	0.0198	0.1088	0.1896	7.30	0.0000
2=No						
DARE Program						
1=Exposed	0.9089	0.0161	0.8704	0.9368	56.30	0.0000
2=Not Exposed	0.8555	0.0198	0.8104	0.8912	43.19	0.0000

Exhibit 19 indicates that the conditional marginal proportions (model-adjusted risks for smoking initiation) are 9.11% for those exposed, and 14.45% for those not exposed. Both are just slightly smaller than the unadjusted risks produced by PROC DESCRIPT. The 95% confidence limits on the marginals just slightly overlap.

Exhibit 20. Model-Adjusted Risk Ratios (R=Independent, SEMETHOD=Zeger)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Zeger-Liang, 1986)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

Initiation of Cigarette Use: Conditional Marginal Risk Ratio #1	CONDMARG Risk Ratio	SE	Lower 95% Limit	Upper 95% Limit
1=Yes				
DARE Program				
1=Exposed vs. 2=Not Exposed	0.630	0.135	0.408	0.973
2=No				
DARE Program				
1=Exposed vs. 2=Not Exposed	1.062	0.030	1.004	1.124

Exhibit 20 indicates that the model-adjusted risk ratio for smoking initiation among those exposed vs. not exposed to DARE is 0.63, with a 95% CI that does not contain the null value of 1.0. This is in agreement with the odds ratio of 0.593 and the overall significance level for the DARE effect. The reduction in risk for those exposed to DARE is statistically significant.

R=Independent, SEMETHOD=Model

In the interest of space, we omit the output obtained under working independence using the model-based or naive variance-covariance matrix of the estimated regression coefficients. ***Exhibit 6*** contains the DARE effect obtained under these conditions.

The model-based variance is the $\boldsymbol{M}_0^{-1}$ matrix, or the outside portion of the robust variance estimate, $\boldsymbol{M}_0^{-1} = [\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}]^{-1}$, where $\boldsymbol{D} = \partial\boldsymbol{\pi}_i / \partial\boldsymbol{\beta}$ is the vector of first partial derivatives of the response probabilities, $\boldsymbol{\pi}_i$, with respect to the regression coefficients, $\boldsymbol{\beta}$. In this case, the naive variance estimate is computed as if the independent working correlation assumption were correct. In other words, these are the results that would be obtained if clustering were ignored altogether. Although it is not recommended for analysis of clustered data, we produced results under these assumptions to demonstrate the effects of clustering. We use the R=Independent and SEMETHOD=MODEL option on the PROC statement to obtain the model-based results.

The estimated regression coefficients are the same as previously, but the estimated standard errors using the model-based approach under independence are much smaller than with the robust variance estimator. The effects of DARE (p=0.0069), family situation (p=0.0358), and grade in school (p=0.0317) are all statistically significant. These standard error estimates are overly optimistic (naive), computed as if the data were truly independent. Therefore, these results are <u>not valid</u> for the data at hand. They merely demonstrate the consequences of ignoring the experimental design.

R=Exchangeable, SEMETHOD=Zeger

We also omit the output obtained from the logistic regression model via the GEE model-fitting technique under the assumption of exchangeable working correlations (R=exchangeable) and using a robust variance estimator (SEMETHOD=Zeger). Results are summarized here and in ***Exhibit 2***.

The estimated intracluster correlation is 0.0206. This value is used in estimating the final regression parameters. In this example, the treatment effect (DARE) has become slightly more significant (p=0.0216) under exchangeability, as the parameter estimate (-0.5825) has increased compared to working independence (-0.5225). The variance estimate has also increased, but only slightly. The estimated odds ratio for initiating smoking by Wave 2 is now 0.56 under exchangeability, vs. 0.59 under working independence. Nevertheless, the overall conclusions are qualitatively the same as for independent working correlations with a robust variance estimate.

All of the effects have become slightly more significant under exchangeability compared to working independence with a robust variance estimate. However, this should not be taken as a general result for exchangeability vs. working independence. Studies have shown that modeling the correlations tend to yield greater power for detecting within-cluster covariates (Neuhaus and Segal, 1993; Lipsitz, Fitzmaurice, Orav, and Laird, 1994), such as sex, race, and family status in the current example. Cluster-level covariates such as the DARE effect seem not to benefit as much from modeling the correlation structure.

R=Exchangeable, SEMETHOD=Model

Below are the results from the exchangeable correlations model using the model-based variance-covariance matrix of the estimated regression coefficients (***Exhibit 21***). The model-based variance is the $\boldsymbol{M}_0^{-1}$ matrix, or the outside portion of the robust variance estimate, $\boldsymbol{M}_0^{-1} = [\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}]^{-1}$, where $\boldsymbol{D} = \partial\boldsymbol{\pi}_i / \partial\boldsymbol{\beta}$ is the vector of first partial derivatives of the response probabilities, $\boldsymbol{\pi}_i$, with respect to the regression coefficients, $\boldsymbol{\beta}$. In this case, the model-based variance estimate is computed assuming that the exchangeable working correlation assumption were correct.

Exhibit 21. MULTILOG Code (R=Exchangeable, SEMETHOD=Model)

```
PROC MULTILOG DATA=one FILETYPE=SAS SEMETHOD=MODEL R=EXCHANGEABLE;
  NEST _ONE_ SCHOOL;
  WEIGHT _ONE_;

  CLASS DARE FIFTH SEX RACE OTHFAM AREA INTCIG12;
  MODEL INTCIG12 = DARE FIFTH SEX RACE OTHFAM AREA / CUMLOGIT;
  CONDMARG DARE / adjrr;

  SETENV LABWIDTH=28 COLWIDTH=7 DECWIDTH=4 COLSPCE=2 TOPMGN=0;
  PRINT  / betas=default risk=default tests=default rhos=default
         orfmt=f5.3 loworfmt=f9.3 uporfmt=f9.3
         t_betafmt=f6.2 waldffmt=f6.2 dffmt=f7.0;

  SETENV LABWIDTH=22 COLWIDTH=6 DECWIDTH=4 COLSPCE=1 TOPMGN=0;
  PRINT / COND_MRG=default condmrgfmt=f11.4 t_cndmrgfmt=f8.2 p_cndmrgfmt=f8.4;

  SETENV LABWIDTH=40 DECWIDTH=3 COLSPCE=1 TOPMGN=0;
  PRINT / CONDRISK=default cond_rrfmt=f8.3;

  RLABEL DARE="DARE Program";
  RLABEL INTCIG12="Initiation of Cigarette Use";
  RLABEL FIFTH="Grade in School";
  RLABEL OTHFAM="Family Situation";
  RFORMAT DARE dare.;
  RFORMAT INTCIG12 yesno.;
  RFORMAT FIFTH grade.;
  RFORMAT SEX sex.;
  RFORMAT RACE race.;
  RFORMAT OTHFAM family.;
  RFORMAT AREA area.;
  RTITLE "MULTILOG Logistic Regression Model for the DARE Evaluation Study"
         "Ennett, et al, 1994";
```

Exhibit 22. First Page of MULTILOG Output (R=Exchangeable, SEMETHOD=Model)

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute    February 2011
                               Release 11.0.0


DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method,
Assuming a With Replacement (WR) Design
    Sample Weight: _ONE_
    Stratification Variables(s): _ONE_
    Primary Sampling Unit: SCHOOL
    Cluster Identification Variables: _ONE_    SCHOOL


Independence parameters have converged in 5 iterations

Step 1 parameters have converged in 6 iterations.

Number of observations read       :   1525    Weighted count:     1525
Observations used in the analysis :   1188    Weighted count:     1188
Denominator degrees of freedom    :     35


Maximum number of estimable parameters for the model is 10

File ONE contains   36 Clusters
  36 clusters were used to fit the model
Maximum cluster size is 123 records
Minimum cluster size is   8 records


Sample and Population Counts for Response Variable INTCIG12
Based on observations used in the analysis
  1=Yes:  Sample Count      145    Population Count       145
  2=No :  Sample Count     1043    Population Count      1043
```

By default, SUDAAN fits the one-step GEE estimates (Lipsitz et al., 1994). Here, we see that the independence betas (the starting estimates for GEE exchangeable) have converged in five iterations, and the Step 1 GEE parameter estimates (under exchangeable working correlations) have converged in six iterations (***Exhibit 26***).

Exhibit 23. Estimated Intracluster Correlation (R=Exchangeable, SEMETHOD=Model)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Model-Based (Naive)
Working Correlations: Exchangeable
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994

Correlation Matrix

-------------------------------------
Initiation of Cigarette Use   Initiation of Cigarette Use
                                1=Yes
-------------------------------------
1=Yes                          0.0206
-------------------------------------
```

Exhibit 23 presents the estimated correlation matrix, which has only one parameter because the response is binary. We see that the estimated intracluster correlation is 0.0206. This value will be used in estimating the final regression parameters.

Exhibit 24. Regression Coefficient Estimates (R=Exchangeable, SEMETHOD=Model)

Variance Estimation Method: Taylor Series (WR)
SE Method: Model-Based (Naive)
Working Correlations: Exchangeable
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994

INTCIG12 (cum-logit), Independent Variables And Effects	Beta Coeff.	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0	P-value T-Test B=0
INTCIG12 (cum-logit)						
Intercept 1: 1=Yes	-1.8802	0.3699	-2.6311	-1.1293	-5.08	0.0000
DARE Program						
1=Exposed	-0.5825	0.2433	-1.0764	-0.0886	-2.39	0.0221
2=Not Exposed	0.0000	0.0000	0.0000	0.0000	.	.
Grade in School						
1=5th Grade	-0.4629	0.2703	-1.0117	0.0859	-1.71	0.0957
2=6th Grade	0.0000	0.0000	0.0000	0.0000	.	.
SEX						
1=Male	0.0876	0.1813	-0.2804	0.4556	0.48	0.6320
2=Female	0.0000	0.0000	0.0000	0.0000	.	.
RACE						
1=Black	0.5088	0.3042	-0.1087	1.1263	1.67	0.1033
2=Hispanic	0.2778	0.3795	-0.4926	1.0482	0.73	0.4690
3=Other	0.5180	0.2876	-0.0658	1.1019	1.80	0.0803
4=White	0.0000	0.0000	0.0000	0.0000	.	.
Family Situation						
1=Non-Traditional	0.4366	0.1937	0.0433	0.8299	2.25	0.0306
2=Traditional	0.0000	0.0000	0.0000	0.0000	.	.
AREA						
1=Rural	-0.0676	0.3759	-0.8308	0.6955	-0.18	0.8582
2=Suburban	-0.2616	0.3447	-0.9614	0.4381	-0.76	0.4529
3=Urban	0.0000	0.0000	0.0000	0.0000	.	.

Exhibit 24 presents the estimated regression coefficients computed under exchangeability and the standard errors as if the exchangeable working assumption were correct. The standard errors are roughly the same as with the robust variance estimator for these data.

Exhibit 25. ANOVA Table (R=Exchangeable, SEMETHOD=Model)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Model-Based (Naive)
Working Correlations: Exchangeable
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	10	26.22	0.0000
MODEL MINUS INTERCEPT	9	2.48	0.0260
DARE	1	5.73	0.0221
FIFTH	1	2.93	0.0957
SEX	1	0.23	0.6320
RACE	3	1.35	0.2725
OTHFAM	1	5.08	0.0306
AREA	2	0.32	0.7270

Exhibit 25 presents the main effect tests computed under exchangeability, using the model-based variance approach. Results are essentially the same as exchangeability with a robust variance estimator, both of which are slightly more significant than working independence with a robust variance estimator.

Exhibit 26. Default Odds Ratios (R=Exchangeable, SEMETHOD=Model)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Model-Based (Naive)
Working Correlations: Exchangeable
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

INTCIG12 (cum-logit), Independent Variables and Effects	Odds Ratio	Lower 95% Limit OR	Upper 95% Limit OR
INTCIG12 (cum-logit)			
Intercept 1: 1=Yes	0.153	0.072	0.323
DARE Program			
1=Exposed	0.559	0.341	0.915
2=Not Exposed	1.000	1.000	1.000
Grade in School			
1=5th Grade	0.629	0.364	1.090
2=6th Grade	1.000	1.000	1.000
SEX			
1=Male	1.092	0.755	1.577
2=Female	1.000	1.000	1.000
RACE			
1=Black	1.663	0.897	3.084
2=Hispanic	1.320	0.611	2.852
3=Other	1.679	0.936	3.010
4=White	1.000	1.000	1.000
Family Situation			
1=Non-Traditional	1.547	1.044	2.293
2=Traditional	1.000	1.000	1.000
AREA			
1=Rural	0.935	0.436	2.005
2=Suburban	0.770	0.382	1.550
3=Urban	1.000	1.000	1.000

The estimated odds ratio for DARE is 0.559 (***Exhibit 26***), compared to 0.593 under working independence. Modelling the correlations yielded slightly more pronounced effects.

Exhibit 27. Conditional Marginal Proportions (R=Exchangeable, SEMETHOD=Model)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Model-Based (Naive)
Working Correlations: Exchangeable
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

Initiation of Cigarette Use						
Conditional Marginal #1	Conditional Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0	P-value
1=Yes						
DARE Program						
1=Exposed	0.0870	0.0152	0.0607	0.1233	5.71	0.0000
2=Not Exposed	0.1458	0.0202	0.1094	0.1918	7.22	0.0000
2=No						
DARE Program						
1=Exposed	0.9130	0.0152	0.8767	0.9393	59.94	0.0000
2=Not Exposed	0.8542	0.0202	0.8082	0.8906	42.27	0.0000

The conditional marginal proportions (model-adjusted risks for smoking initiation) are 8.7% for those exposed, and 14.6% for those not exposed (***Exhibit 27***). The 95% confidence intervals on the marginals are only slightly overlapping. These results are very similar to those produced under independence with a robust variance estimator, yet there is a slightly larger effect under exchangeability.

Exhibit 28 indicates that the model-adjusted risk ratio=0.597, with a 95% CI that does not contain the null value of 1.0. This is in agreement with the results produced under independence with a robust variance estimator, yet slightly more significant here.

Exhibit 28. Model-Adjusted Risk Ratios (R=Exchangeable, SEMETHOD=Model)

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Model-Based (Naive)
Working Correlations: Exchangeable
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Ennett, et al, 1994
```

Initiation of Cigarette Use Conditional Marginal Risk Ratio #1	CONDMARG Risk Ratio	SE	Lower 95% Limit	Upper 95% Limit
1=Yes				
DARE Program				
1=Exposed vs. 2=Not Exposed	0.597	0.129	0.385	0.926
2=No				
DARE Program				
1=Exposed vs. 2=Not Exposed	1.069	0.030	1.010	1.132