Analyzing Survey Data Using SUDAAN®
Release 7.5

by

Gayle S. Bieler
gbmac@rti.org
Research Triangle Institute

and

Rick L. Williams
willy@rti.org
Research Triangle Institute

Prepared for:
Joint Statistical Meetings
Continuing Education Workshop
August, 1997

Analyzing Survey Data Using SUDAAN Release 7.5 was written by **Gayle S. Bieler and Rick L. Williams**

Copyright 1997 by Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709-2194

All rights reserved. No part of this publication may be reproduced or transmitted by any means without permission from the publisher.

SUDAAN and RTI are trademarks of the Research Triangle Institute.
SAS is a trademark of SAS Institute, Inc.
SPSS is a trademark of SPSS, Inc.

How to contact us:

Voice: **919-541-6236**
Fax: **919-541-7431**
Email: **sudaan@rti.org**

Visit SUDAAN on the Web!

www.rti.org/patents/sudaan/sudaan.html

SUDAAN

Software for the Statistical Analysis of Correlated Data

Analyzing Survey Data Using SUDAAN® Release 7.5

by

Gayle S. Bieler
gbmac@rti.org

and

Rick L. Williams
willy@rti.org
Research Triangle Institute

Prepared for:
Joint Statistical Meetings
Continuing Education Workshop
August, 1997

Research Triangle Institute
PO Box 12194
Research Triangle Park, NC 27709-2194

Analyzing Survey Data Using SUDAAN Release 7.5

Table of Contents

Analyzing Survey Data Using SUDAAN® Release 7.5

ABSTRACT

In the social sciences and public health, researchers often analyze survey data which were collected via a complex sampling design. Such survey designs often include stratification and cluster sampling (*e.g.*, sampling by geographic clusters) in one or more stages, where the clusters may be sampled with unequal probabilities. Such designs complicate the statistical analysis since the observations are not independent and identically distributed (*iid*). Failure to account for the design in the statistical analysis typically result in underestimated standard errors and false positive test results.

Unlike standard statistical packages, SUDAAN is specifically designed to handle non-*iid* observations drawn from finite populations. SUDAAN offers a powerful set of analytic tools for linear regression, logistic regression, multinomial logistic regression, proportional hazards modelling, and descriptive data analysis.

This seminar will review the statistical methods used in SUDAAN and demonstrate its use via a series of examples from the social sciences and public health. The basic concept throughout all SUDAAN procedures is to use consistent variance estimators for statistics derived from complex samples (*e.g.*, means, proportions, odds ratios, regression coefficients), without imposing strict distributional assumptions, and treating the intracluster correlation as a nuisance parameter. ***SUDAAN is currently the only statistical package to offer three well-known methods for variance estimation in sample surveys: Taylor linearization, BRR, and Jackknife.***

This workshop will highlight many of the new features in SUDAAN Release 7.5 that are of particular interest to survey researchers, including**: 1)** BRR and Jackknife variance estimation for descriptive statistics and regression modelling; **2)** Four choices for computing ***design effects***; **3)** ***User-friendly contrast statements*** and a ***reference level statement*** for specifying the reference cells of categorical covariates in all regression procedures; **4)** ***Least squares means*** estimation for linear regression; **5)** a more useful R^2 statistic based on the log-likelihood for logistic regression; and **6)** better compatibility with well-known software packages (***SAS-Callable versions*** for SUN Solaris and Win 95, and ***reading SPSS datasets***). Attendees should be familiar with the basics of survey sampling and analysis, as well as fitting linear and non-linear regression models.

About Sample Surveys

What is a Sample Survey?

- A study involving a subset (or sample) of individuals selected from a larger population with known probabilities of selection
- Measurements are aggregated over all sample members to obtain summary statistics (*e.g.*, means, proportions, totals, or ratios) for the sample
- Extrapolations made to the entire population (estimates of population parameters)

What is a Census?

- All individuals in a population are selected for measurement (summary stats are not extrapolations)

Main Advantages of Sampling:

- Reduced Cost
- Greater Speed

Many Surveys are Purely Descriptive:

- Estimation of summary statistics often the primary objective; hypothesis testing a secondary objective

Characteristics Of Complex Sample Surveys

Target Population

Entire set of individuals to which findings are to be extrapolated
Individual members of the population whose characteristics are to be measured are called ***population elements***

Example:

Select a nationally representative sample of students from the US population

Sample Design:

Stratified, Multi-Stage Nested Design

1) Divide the country into 4 ***regions*** (***strata***)

2) Obtain a comprehensive list of schools in each region; Select a sample of ***schools*** from each region, according to a known probability sampling scheme, such as:

   - simple random sampling,
   - probability proportional to size sampling (PPS),
   - certainty sampling (probability of selection = 1)

3) Obtain a comprehensive list of ***students*** within each sample school;

   Select a sample of ***students*** from each school according to a known probability sampling scheme (as above), or select all students from the school (probability of selection = 1)

Characteristics Of Complex Sample Surveys

Stratification

- Selection of sampling units (population elements) from mutually exclusive and exhaustive subpopulations

- **STUDENT SAMPLE: Strata = REGION**

  Regions were ***not*** randomly selected; they were chosen in advance

  Independent samples of schools chosen within each stratum

- Different sampling methods can be used in different strata

- Obtain an estimate for the population as a whole by aggregating the individual stratum estimates over all strata

- Can ***reduce variance*** of sample statistics (e.g., average GPA, average height) if strata are chosen efficiently (i.e., if strata are homogeneous wrt the variable of interest).

Characteristics Of Complex Sample Surveys

Clustering

Problems with direct element sampling if:

- There exists no sampling frame for the population elements (*e.g.*, no master list of students in US from which to select a sample)

- The population elements are scattered over a wide area in which case direct element sampling will result in a scattered sample

  - field costs prohibitive

Solution:

- Use cluster sampling

  Population elements are aggregated into larger units (***clusters***) for which complete lists are available

  e.g., schools

- Use multistage designs

Characteristics of Complex Sample Surveys

Clustering

- Subunits (students) are selected into the sample from clusters or primary sampling units (schools)

  Student Sample:

  Primary Sampling Unit (*PSU* or *Cluster*) = *School*
  Students clustered within schools;

- Usually positive correlation within clusters (*i.e.*, students within schools are more alike than across schools, so they tend to respond similarly)

- Variance of sample statistics (e.g., average GPA) is typically ***increased*** under cluster sampling

  How Much ?

Characteristics of Complex Sample Surveys

Clustering

- ***Design Effect***

  Describes the change in variance of an estimated statistic due to clustering

  Estimated as the ratio of variance under the cluster design vs. a simple random sample of the same size (*i.e.,* independence), or via an analytic expression:

$$Design\ \ Effect \ = \ \frac{V(\hat{\theta})_{CLUSTER}}{V(\hat{\theta})_{SRS}} \ = \ 1 \ + \ \rho(m-1) \ ,$$

  m = average cluster size (students)
  ρ = intra-cluster correlation coefficient
  (measure of association within the cluster)

- If $\rho = 0$:

  No correlation (DEFF = 1)

  If $\rho = 1$:

  Perfect positive correlation (*e.g.*, everyone responds the same)
  DEFF = m (cluster size)

  If $0 < \rho < 1$:
  Some degree of correlation (units respond similarly)
  $1 <$ DEFF $< m$

Characteristics of Complex Sample Surveys

Clustering

- For a sample of a given size, as the cluster size and the intracluster correlation increase, the variance is also ***increased***.

- Another way to think about clustering:

    Loss of precision
    Reduction in effective sample size

  Effective sample size < number of observations (students)
  > number of clusters (schools)

Characteristics of Complex Sample Surveys

Multi-Stage Sampling

- Cluster sampling in 2 or more stages:

- Selection Order:

  Stage 1: Counties
  Stage 2: Schools from sampled counties
  Stage 3: Students from sampled schools

  or

  Stage 1: Schools (stratified by region)
  Stage 2: Homeroom classes (stratified by grade)
  Stage 3: Students.

Characteristics of Complex Sample Surveys

Unequal Weighting

- Relates to ***Probability*** or ***Population-Based Sampling***

  - Every element in the target population has a known, non-zero probability of being included in the sample

- ***Unequal weighting*** results when sample members (*e.g.*, students) selected with unequal probabilities

  - Oversampling certain subpopulations, such as the elderly, the poor, Hispanics, or Native Americans.

- Each sample member has a sampling weight associated with their data

  Sampling weight = inverse of selection probability

  Refers to number of individuals in target population that the sample member represents

  Weights needed for unbiased estimation of population parameters (findings are then generalizable to a finite population of interest).

- Downside: Variability in sampling weights can lead to inefficiency, meaning loss of power and wider confidence intervals.

- Variance of sample statistics usually ***increased*** if weights are highly variable.

Characteristics of Complex Sample Surveys

Nonlinear Statistics

- Most survey statistics are ***not*** simple linear functions of the data, but rather ratios of random variables

***Linear Statistic*:** the weighted total

$$x_w = \sum_{i=1}^{n} w_i x_i$$

***Nonlinear Statistic*:** the weighted mean, proportion, etc.

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad w_i = \frac{1}{p_i} = \textit{sampling weight for sample member } i$$

$$x_i = \textit{outcome of interest for sample member } i$$

- $\bar{x}_w$ is a weighted mean if the response variable is continuous, or a weighted proportion if x_i is coded 0 (characteristic absent) and 1 (characteristic present)

- The denominator $\sum w_i$ is not a fixed quantity but rather an *estimate* of the population size, which varies from sample to sample when the weights are unequal.

- Non-standard techniques required for variance estimation: SUDAAN offers ***Taylor series linearization*** and ***replication methods (BRR and Jackknife)***

Characteristics of Complex Sample Surveys

Without Replacement Sampling

- Units selected into the sample do not have another chance of being selected

- In practice, almost all sampling is done without replacement, but can often be ignored in the analysis

- If you account for it in analysis, the variance of sample statistics is ***decreased*** when the sampling fractions (*e.g.*, proportion of schools selected from each region) become large

- Sampling fraction:

$$\frac{n}{N} = \frac{number\ units\ selected\ into\ sample}{number\ units\ in\ population}$$

- In other words, the more you know about a population, the smaller the variance of sample statistics

- ***Why is it ignored most of the time?***
  Accounting for without-replacement sampling makes variance estimation slightly more complicated, since you must know the sampling fractions within each of the first-stage strata

  There is little efficiency to be gained when sampling fractions are small.

Applications in Epidemiologic Studies

Longitudinal Studies
Repeated Measures Studies

Multiple events, such as hospital visits or illness episodes, are observed over time on each subject.

Example 1:
Relationship between MDI (Mental Development Index) measurements and umbilical cord blood lead levels in children (Waternaux, et al, *JASA*, 1989)

MDI measurements recorded at 6, 12, and 18 months of age for each child

Example 2:
Logistic regression of the propensity of daily asthma attacks on the average daily level of total suspended particulates in the air (Korn and Whittemore, 1979, *Biometrics*)

Daily asthma measurements and other time-dependent covariates recorded on each person in a sample of adults and children (up to 34 weeks of daily measurements)

Applications in Behavioral Research

Examples:

- School-based evaluations of substance-abuse prevention programs in the student population (observations are on students nested within schools)

- Evaluation of Project DARE (Ennett *et al*, 1994, *Addictive Behaviors*; Norton, *et al.*, 1996, *Journal of Consulting and Clinical Psychology*)

Multi-Stage Sample Surveys

- Data are obtained via a complex survey design (Cluster sampling in 1 or more stages; clusters may be sampled with differing probabilities)

- Practical advantages to multi-stage design (e.g., sampling by geographic clusters):

  – Not always feasible to enumerate the population of interest (sample frame)
  – Reduces cost of data collection (travel)

- Design-based methods of analysis:

  – Weighting of the data for unbiased estimates
  – ***Linearization*** and ***replication methods*** to estimate variances

- Examples: NHANES, NHIS, BRFSS, NHSDA

Sudaan vs. Other Software (SAS®, SPSS®, ...)

SAS, SPSS, etc		**SUDAAN**
Simple random sampling; Infinite populations	SAMPLE SELECTION ASSUMPTIONS	Complex probability sampling schemes; finite populations
Known probability distributions (normal, binomial)	DISTRIBUTIONAL ASSUMPTIONS	No strict distributional assumptions
Linear statistics ***only***	RANDOM VARIABLE ASSUMPTIONS	Functions of linear statistics

RESULT: ***SAS*** yields unbiased point estimates if you include appropriate weights, but variance estimates ***wrong*** (usually ***underestimated***) due to clustering.

Test statistics have inflated Type I error rates (reject null hypothesis more often than nominally specified, *i.e.*, false positives)

SUDAAN yields consistent variance estimates for sample statistics (*e.g.*, means, totals, proportions, ratios, regression coefficients) needed for unbiased inference.

Why SUDAAN?

An Example

WIC Mothers and Infants:	Two-stage clustered design Strata = Region PSU = WIC local agencies
Sample Size:	953
Population Size:	Approximately 506,000 WIC participants
Outcome of Interest:	Initiated breastfeeding
ESTIMATE:	Percentage breastfed their infant
COMPARISON DOMAINS:	Race groups (white vs. non-whites)

(Results Follow)

Why SUDAAN?

An Example

SAS Results

Association Between Breast-feeding and Mother's Race

Sampling Weights Sum to Population Size

```
                TABLE OF BFEED BY MOMRACE

           BFEED (Breastfeeding Initiation)
                     MOMRACE (Mother Race)

           Frequency|
           Col Pct  |White   |Other   |  Total
           ---------------------------|
           Yes      | 129929 | 141964 | 271893
                    |  51.04 |  56.40 |
           ---------------------------|
           No       | 124638 | 109748 | 234386
                    |  48.96 |  43.60 |
           ---------------------------|
           Total      254567   251712   506279


        STATISTICS FOR TABLE OF BFEED BY MOMRACE

    Statistic              DF      Value        Prob
    ----------------------------------------------
    Chi-Square              1   1462.544       0.001

    Sample Size = 506279
```

Why SUDAAN?

An Example

SAS Results

Association Between Breast-feeding and Mother's Race

Weights Normalized to Sum to Sample Size:
NORMWGT = WEIGHT * (953 / 506,279)

```
                     Normalized Weights

                TABLE OF BFEED BY MOMRACE

           BFEED (Breastfeeding Initiation)
                    MOMRACE (Mother Race)

           Frequency|
           Col Pct  |White   |Other   |  Total
           ---------------------------|
           Yes      | 244.57 | 267.23 |  511.8
                    |  51.04 |  56.40 |
           ---------------------------|
           No       | 234.61 | 206.59 |  441.2
                    |  48.96 |  43.60 |
           ---------------------------|
           Total     479.187  473.813      953


        STATISTICS FOR TABLE OF BFEED BY MOMRACE

    Statistic              DF      Value        Prob
    -----------------------------------------------
    Chi-Square              1      2.753       0.097

    Sample Size = 953
```

Why SUDAAN?

An Example

SUDAAN Results

Association Between Breast-feeding and Mother's Race

Weights Sum to Population Size

```
Date: 07-17-97           Research Triangle Institute           Page  : 2
Time: 16:13:03             The CROSSTAB Procedure              Table : 1


Number of observations read    :    953    Weighted count :   506279
Denominator degrees of freedom :     21


Variance Estimation Method: Taylor Series (WR)

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

-----------------------------------------------------------------------------
| Breastfeeding     |                  | Mother Race
| Initiation        |                  | Total       | White       | Non-White  |
-----------------------------------------------------------------------------
| Total             | Sample Size      |         953 |         480 |         473 |
|                   | Population       |      506279 |      254567 |      251712 |
|                   | Column Percent   |      100.00 |      100.00 |      100.00 |
|                   | Std Error        |        0.00 |        0.00 |        0.00 |
|                   | Design Effect    |           . |           . |           . |
-----------------------------------------------------------------------------
| Yes               | Sample Size      |         522 |         249 |         273 |
|                   | Population       |      271893 |      129929 |      141964 |
|                   | Column Percent   |       53.70 |       51.04 |       56.40 |
|                   | Std Error        |        3.10 |        3.18 |        4.99 |
|                   | Design Effect    |        3.74 |        1.97 |        4.90 |
-----------------------------------------------------------------------------
| No                | Sample Size      |         431 |         231 |         200 |
|                   | Population       |      234386 |      124638 |      109748 |
|                   | Column Percent   |       46.30 |       48.96 |       43.60 |
|                   | Std Error        |        3.10 |        3.18 |        4.99 |
|                   | Design Effect    |        3.58 |        1.91 |        4.64 |
-----------------------------------------------------------------------------
```

Why SUDAAN?

An Example

SUDAAN Results

Association Between Breast-feeding and Mother's Race

Weights Sum to Population Size

```
Date: 07-17-97              Research Triangle Institute            Page  : 1
Time: 16:13:03                The CROSSTAB Procedure               Table : 1


Variance Estimation Method: Taylor Series (WR)

Chi Square Test of Independence for Breastfeeding Initiation
  and Mother Race

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

-------------------------------------------------
|                 |                  |
|                 |                  |          |
-------------------------------------------------
|                 |                  |          |
|                 | Chi-Square       |   0.9403 |
|                 | DF               |        1 |
|                 | P-Value          |   0.3432 |
-------------------------------------------------
```

Why SUDAAN?

An Example

SUMMARY OF RESULTS

Package	Method	%Breastfeed (SE): **White**	%Breastfeed (SE): **Non-White**	Chi-Square	P-value
SAS	Weighted	51.04 (0.10)	56.40 (0.10)	1462.5	0.001
	Weights Normalized (Sum to Sample Size)	51.04 (2.28)	56.40 (2.28)	2.75	0.097
SUDAAN	Weighted	51.04 (3.18)	56.40 (4.99)	0.94 *	0.343

NOTE:
SAS standard errors are calculated as $\sqrt{p\,(1-p)\,/\,n}$, where n is the sum of the weights, and p is the proportion breastfeeding.

Why Did We Bother Developing SUDAAN?

Intra-Cluster Correlation

- Potential for clustermates to respond similarly (genetic and environmental influences)

- Experimental units from the same cluster are not statistically independent

- Usually results in ***overdispersion***, or extra-variation in the responses beyond what would be expected under independence

- Other standard statistical packages (*e.g.*, SAS®, SPSS®) do not uniformly address the correlated data problem in all analytical procedures

  SUDAAN uses correlated data methods for:

  - Regression modelling
  - Estimating and analyzing:
    Means, medians, percentages, percentiles, odds ratios and relative risks, and ratios of random variables
  - Chi-square tests in contingency tables
  - Cochran-Mantel-Haenszel tests in contingency tables

Multivariate Responses (Clustered Data)

Notation

$$i = PSU \text{ or cluster}$$
$$= 1, \ldots, n$$
$$j = \text{observation within the cluster}$$
$$= 1, \ldots, m_i$$

Data

$$(y_{ij}, \boldsymbol{x}_{ij}) , \quad j = 1, \ldots, m_i$$
$$i = 1, \ldots, n$$
$$N = \sum_i m_i = \text{total sample size}$$

Responses

$$\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{im_i})$$

Covariates

$$\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})$$

This is the clustered data situation covered by SUDAAN

Assumptions: Independence Vs. Clustered Data

Independence

$$
\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix} \qquad \boldsymbol{V}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I_N} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}
$$

Observations independent, constant variance

Clustered Data (SUDAAN):

$$
\boldsymbol{Y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1m_1} \\ \vdots \\ y_{n1} \\ \vdots \\ y_{nm_n} \end{bmatrix}
$$

n clusters of m_i observations $\left(N = \sum_{i=1}^{n} m_i\right)$

Unequal observations per cluster $= m_i$

Example: n litters with m_i pups per litter

Assumptions: Independence Vs. Clustered Data

Clustered Data (SUDAAN):

$$\boldsymbol{V(Y)} = \begin{bmatrix} V_1 & 0 & 0 & \cdots & 0 \\ 0 & V_2 & 0 & \cdots & 0 \\ 0 & 0 & V_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & V_n \end{bmatrix}$$

Cluster − Correlated Data

Block − Diagonal by Cluster

V_i is an m_i x m_i matrix

$$\boldsymbol{V_i} = \begin{bmatrix} \sigma^2_{(i)1} & \sigma_{(i)12} & \sigma_{(i)13} & \cdots & \sigma_{(i)1m} \\ \sigma_{(i)21} & \sigma^2_{(i)2} & \sigma_{(i)23} & \cdots & \sigma_{(i)2m} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \sigma_{(i)m1} & \sigma_{(i)m2} & \sigma_{(i)m3} & \cdots & \sigma^2_{(i)m} \end{bmatrix}$$

- V_i is an m_i x m_i variance covariance matrix of observations in the *i*-th cluster
- ***No assumptions on structure*** of V_i (could be unstructured, multi-level, AR(1), exchangeable, etc.)
- Observations independent between clusters, completely arbitrary correlation structure within clusters

Independence Vs. Clustered Data: Fitting Linear Regression Models

Standard Situation: Linear Regression

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_N \end{bmatrix} \qquad \begin{aligned} &E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} \\ &\boldsymbol{V}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I_N} \end{aligned}$$

Independent obs, constant variance

Standard Solution to Normal Equations:

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$Var(\boldsymbol{b}) = \hat{\sigma}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$$ $\hat{\sigma}^2$ = Mean Square Error

This variance formula only holds when: $V(\boldsymbol{Y}) = \sigma^2 \boldsymbol{I_N}$

Independence Vs. Clustered Data: Fitting Linear Regression Models

How is SUDAAN different?

$$\boldsymbol{V}(\boldsymbol{Y}) = \boldsymbol{V_Y} = \begin{bmatrix} V_1 & 0 & 0 & \cdots & 0 \\ 0 & V_2 & 0 & \cdots & 0 \\ 0 & 0 & V_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & V_n \end{bmatrix}$$

Cluster−Correlated Data

Block−Diagonal by Cluster

V_i is an m_i x m_i matrix

$$\boldsymbol{b} = (\boldsymbol{X}' \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{Y}$$

Use ***robust variance formula*** to estimate:

$$Var(\boldsymbol{b}) = \boldsymbol{V_b}$$

Estimates each element separately

KEY POINT:

$$V_{\boldsymbol{b}} \neq \hat{\sigma}^2 (X'X)^{-1}$$

due to cluster−correlated data

Independence Vs. Clustered Data Fitting Linear Regression Models

Null Hypothesis:

$$H_0:\ \boldsymbol{C\beta} = \mathbf{0}$$

C is a contrast matrix of rank r

General Form for Test Statistic:

$$Q = (\boldsymbol{Cb})' \left[\boldsymbol{C}\ Var(\boldsymbol{b})\ \boldsymbol{C}'\right]^{-1} (\boldsymbol{Cb})$$

Standard Situation

$$Q = (\boldsymbol{Cb})' \left[\hat{\sigma}^2\, \boldsymbol{C}\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right]^{-1} (\boldsymbol{Cb})$$

$$= \frac{r \cdot MS_{H_0}}{MS_{error}} \sim r\, F_{r,\, N-r}$$

Standard computing formula used by most software packages

SUDAAN Test Statistic:

$$Q = (\boldsymbol{Cb})' \left[\boldsymbol{C}\boldsymbol{V_b}\boldsymbol{C}'\right]^{-1} (\boldsymbol{Cb})$$

Does not reduce to any simple computing formula

SUDAAN Software Package

Software for Statistical Analysis of Correlated Data

- Single program, written in the *C* language, consisting of a family of statistical procedures

- As easy to use as SAS!

  - Uses a SAS-like interface
  – Accepts SAS data sets as input

- Two Modes of Operation:

  1) SAS-Callable
     (Win 95, SUN/Solaris, VAX/VMS, IBM/MVS)

  2) Stand-Alone
     (many platforms, including Windows)

- SPSS Users: Release 7.5 reads SPSS files

SUDAAN Procedures

DESCRIPTIVE PROCEDURES

CROSSTAB

Computes frequencies, percentage distributions, odds ratios, relative risks, and their standard errors (or confidence intervals) for user-specified cross-tabulations, as well as chi-square tests of independence and the Cochran-Mantel-Haenszel chi-square test for stratified two-way tables.

DESCRIPT

Computes estimates of means, totals, proportions, percentages, geometric means, quantiles, and their standard errors; also computes standardized estimates and tests of single degree-of-freedom contrasts among levels of a categorical variable.

RATIO

Computes estimates and standard errors of generalized ratios of the form $\Sigma y / \Sigma x$, where x and y are observed variables; also computes standardized estimates and tests single-degree-of-freedom contrasts among levels of a categorical variable.

REGRESSION PROCEDURES

REGRESS

Fits linear regression models and performs hypothesis tests concerning the model parameters. Uses ***GEE*** to efficiently estimate regression parameters, with robust and model-based variance estimation.

LOGISTIC

Fits logistic regression models to binary data and computes hypothesis tests for model parameters; also estimates odds ratios and their 95% confidence intervals for each model parameter.

MULTILOG

Fits logistic and multinomial logistic regression models to ordinal and nominal categorical data and computes hypothesis tests for model parameters; estimates odds ratios and their 95% confidence intervals for each model parameter; uses ***GEE*** to efficiently estimate regression parameters, with robust and model-based variance estimation.

SURVIVAL

Fits discrete and continuous proportional hazards models to failure time data; also estimates hazard ratios and their 95% confidence intervals for each model parameter.

Elements of a SUDAAN Procedure

PROC MULTILOG **DATA** = name
DESIGN = WR | WOR | STRWR | STRWOR |UNEQWOR
JK | BRR ;

WEIGHT *variable*;
REPWGT variable(s);

NEST Strata PSU ;

Primary Sampling Unit
or...
Cluster in experimental designs

For Regression Modelling:

MODEL *dependent = independent* ;

DRUGUSE = AGE SEX RACE ;

For Descriptive Statistics:

VAR *response_variables* ;

TABLE *categorical effects* (*e.g.*, RACE) ;

Enhancements to SUDAAN Release 7.5

Replication Methods for Robust Variance Estimation

- Jackknife
- Balanced Repeated Replication (BRR)

Enhancements of GEE Capabilities

- Exchangeable correlations in linear regression (as already in logistic and multinomial logistic since Release 7.0)
- Robust (default) and model-based variances in GEE applications

Regression Enhancements

- REFLEVEL statement to change the reference level for categorical covariates
- User-friendly contrast statement (EFFECTS) for testing simultaneous regression effects, simple effects in interaction models, and more
- R-square (Cox and Snell, 1989) in logistic regression
- Least Squares Means (LSMEANS) statement in linear regression
- MULTILOG Procedure for multinomial logistic regression (7.0)

SAS-Callable Platforms

- Windows 95
- SUN/Solaris

Now reads SPSS files (in addition to SAS and ASCII)

Three Variance Estimation Methods in SUDAAN

Basic Concept Behind All

1) Use ***consistent estimators*** of the parameters

   e.g., Means, Proportions, Percentages, Odds Ratios, Regression Coefficients

   Can even estimate the correlation structure and improve the efficiency of β

   Intracluster correlation treated as a nuisance parameter

2) ***Robust variance estimators*** ensure consistent variance estimates and valid inferences:

   - Taylor linearization / GEE
   - Jackknife (new in Release 7.5)
   - BRR (new in Release 7.5)

   - *Without* imposing strict distributional assumptions about the response of interest

Taylor Linearization Approach

Two-Step Procedure for Variance Estimation:

1) ***Use Taylor series linearization to approximate functions of linear statistics (e.g., ratios of random variables)***

Example: Prevalence of drug use

$$\hat{p} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}\, y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}} = \frac{Estimated\ \ number\ \ drug\ \ users}{Estimated\ \ population\ \ size}$$

Find ***linear approximation*** to this nonlinear statistic (Kendall and Stuart, 1973);
Design-specific variance formulas available for ***linear*** statistics.

Woodruff (1971):

- Equivalent computational procedure using Taylor series ***linearized values***

- Each observational unit gets a linearized value for a particular statistic.

2) ***Compute design-specific variance of the linearized values***

Taylor Linearization Approach

Design-Specific Variance: Choice of Sample Designs

- ***DESIGN=WOR***
  Equal probability without-replacement sampling at ***each stage***
  (finite population corrections)

- ***DESIGN=UNEQWOR***
  Unequal probability without-replacement sampling at ***first stage***
  (Yates-Grundy-Sen variance estimator)

- ***DESIGN=WR***
  With-replacement sampling at first stage
  (this is referred to as the ***between-cluster variance estimator***)

  - ***Most common choice,*** as long as low sampling fractions at first stage

  - Allows for any sample design within each PSU (*e.g.*, additional stages of sampling, with equal or unequal probabilities of selection)

- Stratification allowed with all designs, even if the sample is not clustered

Between-Cluster Variance Estimator (DESIGN=WR)

Goal is to estimate $Var(\hat{\theta})$:

$$\hat{\theta} = F(X,Y) \quad \textit{where } X \textit{ and } Y \textit{ are linear statistics}$$

$$Z_{ij} = \textit{Linearized value of } \hat{\theta} \textit{ for unit}-ij$$

$$= (\partial F_X)x_{ij} + (\partial F_Y)y_{ij}$$

For a proportion, $\hat{p} = \dfrac{Y}{X}$,

$$Z_{ij} = w_i\,(y_{ij} - \hat{p}) \; / \; \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}$$

$$Z_i = \sum_{j=1}^{m_i} Z_{ij} \qquad \textit{PSU Totals}$$

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \qquad \textit{Mean of PSU Totals}$$

$$\boxed{V\hat{a}r(\hat{\theta}) = \frac{n}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

Implicit Linearization Method for Regression Models

Logistic Regression Model:

$$p_{ijk}(\boldsymbol{\beta}) = \Pr(y_{ijk} = 1 \mid \boldsymbol{x}_{ijk}, \boldsymbol{\beta}) = \left[1 + \exp(-\boldsymbol{x}'_{ijk}\boldsymbol{\beta}) \right]^{-1}$$

where:

i = stratum; j = PSU or cluster; k = observation within the cluster

$\boldsymbol{x}_{ijk} = (1, x_{1,ijk}, \ldots, x_{q,ijk})'$ = vector of regression effects (stratum-, PSU-, and observation-specific)

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)'$ = vector of unknown regression coefficients

$$y_{ijk} = \begin{cases} 1, & \text{outcome present} \\ 0, & \text{outcome absent} \end{cases}$$

Overview of Implicit Taylor Linearization Method (using between-cluster variance estimator)

1) Find solutions to weighted pseudo-likelihood equations (identical to SAS PROC LOGISTIC)

2) Application of Taylor linearization for implicitly-defined parameter vectors in conjunction with a ***between-cluster*** variance estimation formula (Binder, 1983)

   Yields consistent estimator for $Var(\hat{\beta})$

Implicit Linearization Method for Regression Models

Maximize the Log-Likelihood

Weighted Score Equations:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial \operatorname{Log} \mathrm{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i \sum_j \sum_k w_{ijk} \boldsymbol{x}'_{ijk} y_{ijk} - \sum_i \sum_j \sum_k w_{ijk} \boldsymbol{x}'_{ijk} p_{ijk}(\boldsymbol{\beta})$$

Solve via iteration: $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0} \Rightarrow \hat{\boldsymbol{\beta}}$

$$\hat{p}_{ijk} = \left[1 + \exp(-\boldsymbol{x}'_{ijk} \hat{\boldsymbol{\beta}}) \right]^{-1}$$

Binomial-based estimates are asymptotically normally distributed and consistent, even under cluster sampling. Standard regression coefficient estimates are robust to violations of model assumptions.

Weighted Sample Information Matrix:

$$\boldsymbol{J} = -\left[\frac{\partial^2 \operatorname{Log} \mathrm{L}(\boldsymbol{\beta})}{\partial \hat{\boldsymbol{\beta}}^2} \right] = \sum_i \sum_j \sum_k \boldsymbol{x}'_{ijk} \boldsymbol{x}_{ijk} w_{ijk} \hat{d}_{ijk} ,$$

$$\text{where} \quad \hat{d}_{ijk} = \hat{p}_{ijk}(1 - \hat{p}_{ijk})$$

Under Cluster Sampling (Intracluster correlation ≠ 0)

$$V\hat{a}r(\hat{\boldsymbol{\beta}}) \neq \boldsymbol{J}^{-1} \quad \text{Not a Consistent Estimator (Biased)}$$

Implicit Linearization Method for Regression Models

Taylor Linearization for Implicitly-Defined Parameter Vectors
(***Robust*** *or* ***Sandwich Estimator***, Binder 1983)

$$V\hat{a}r\,(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{J}^{-1}\right) V\hat{a}r\left[\hat{\boldsymbol{U}}\,(\hat{\boldsymbol{\beta}})\right] \left(\boldsymbol{J}^{-1}\right)'$$

where

$$\hat{\boldsymbol{U}}\,(\hat{\boldsymbol{\beta}}) = \frac{\partial \operatorname{Log} L\,(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\bigg|_{\beta = \hat{\beta}} \quad \text{Estimating Equations (Score Function)}$$

$$\boldsymbol{J} = \frac{\partial \hat{\boldsymbol{U}}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} \quad \text{Sample Information Matrix}$$

Outside Term

J^{-1} is the ***model-based*** (or ***naive***) variance estimate

Inside Term

$V\hat{a}r\left[\hat{\boldsymbol{U}}(\hat{\boldsymbol{\beta}})\right]$ is the ***design-specific variance correction***

Implicit Linearization Method for Regression Models

Estimate $Var\left[\hat{\mathbf{U}}(\hat{\boldsymbol{\beta}})\right]$ Using Between-Cluster Variance:

1) Score equations are simple linear functions of the observations

$$\hat{U}(\hat{\boldsymbol{\beta}}) = \sum_i \sum_j \sum_k \hat{\boldsymbol{U}}(\boldsymbol{Z}_{ijk}\,;\boldsymbol{\beta})$$

Linearized variate vector for the *ijk*-th unit:

$$\boldsymbol{Z}_{ijk} = w_{ijk}\,\boldsymbol{x}'_{ijk}\,(y_{ijk} - \hat{p}_{ijk})$$

2) Compute *between-PSU* within-stratum variance estimate for a vector of linear statistics:

Accumulations of linearized variate vectors at PSU level

$$\boldsymbol{Z}_{ij} = \sum_k \boldsymbol{Z}_{ijk}\;, \qquad k = 1,\ldots,m_{ij}$$

Form ***Between-PSU*** Within-Stratum Mean Square Matrix

$$\boldsymbol{S}_z = \sum_i n_i\,\boldsymbol{S}_{zi}\,, \qquad n_i = \#\ \text{PSU}'\text{s in stratum i}$$

With sample mean squares and cross-products matrix:

$$\boldsymbol{S}_{zi} = \sum_j (\boldsymbol{Z}_{ij} - \overline{\boldsymbol{Z}}_i)(\boldsymbol{Z}_{ij} - \overline{\boldsymbol{Z}}_i)' \,/\, (n_i - 1)$$

$$\overline{\boldsymbol{Z}}_i = \sum_j \boldsymbol{Z}_{ij} \,/\, n_i\ .$$

Implicit Linearization Method for Regression Models

Estimated Cluster Covariance Matrix for $\hat{\boldsymbol{\beta}}$:

$$V\hat{a}r\,(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{J}^{-1}\right) \boldsymbol{S}_z \left(\boldsymbol{J}^{-1}\right)' .$$

Null Hypothesis:

$$H_0 :\ C\boldsymbol{\beta} = \mathbf{0} \quad vs. \quad H_1 :\ C\boldsymbol{\beta} \neq \mathbf{0}$$

C = Contrast Matrix

Wald Test Statistic:

$$\chi^2 = \left[C\hat{\boldsymbol{\beta}}\right]' \left[C\,\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\,C'\right]^{-1} \left[C\hat{\boldsymbol{\beta}}\right]$$

$$\sim \chi^2_c \,, \quad where\ c = rank\ of\ C$$

Small-Sample Modifications to Wald Chi-Square:

- Wald chi-square too liberal when DF associated with the hypothesis is large compared to the DF available for estimating variance of regression coefficients (#clusters-#strata) (Thomas and Rao, 1987)

- Satterthwaite-corrected Chi-Square (Rao and Scott, 1987)
- Adjusted Wald F-statistic (Folsom, 1974; Fellegi, 1980)

Implicit Linearization Method for Regression Models

Implicit Linearization Method Also Used For:

1) Proportional Hazards Model (Cox Regression)
   Binder, 1992

2) Ordinary Linear Regression: parameter vector ***explicitly*** defined

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}$$

Where ***W*** is a diagonal matrix with diagonal elements equal to the sample member weights.

Estimating Equations (Normal Equations):

$$\hat{\mathbf{U}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}$$

$$\mathbf{J} = \frac{\partial \hat{\mathbf{U}}(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

Linearized Variate Vector for $\hat{U}(\hat{\beta})$

$$Z_{ijk} = \left[x_{ijk}'(x_{ijk}\hat{\beta} - y_{ijk})\right]w_{ijk}$$

Jackknife Variance Estimation

Replication Methods for Complex Survey Data

Quenouille (1956): Reducing bias in estimation
Tukey (1958): Approximate confidence intervals

Start With Given Point Estimator:
Descriptive statistics (*e.g.*, means, proportions)
Regression parameter vectors

- Use consistent estimators of location parameters
- Assumes with-replacement sampling of PSUs (same as the ***between-cluster*** estimator)

Jackknife Variance Estimation

Prevalence of drug use in a complex sample survey:

$$\hat{p} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}\, y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}} = \frac{\textit{Estimated number drug users}}{\textit{Estimated population size}}$$

An estimate based on all PSU's ***except the k-th*** is as follows:

$$\hat{p}_{(k)} = \frac{\sum_{i \neq k}^{n} \sum_{j=1}^{m_i} w_{ij}\, y_{ij}}{\sum_{i \neq k}^{n} \sum_{j=1}^{m_i} w_{ij}}$$

Jackknife Variance Estimate for $\hat{p}$**:**

$$\hat{\sigma}^2_{JK} = \frac{n-1}{n} \sum_{k=1}^{n} \left[\hat{p}_{(k)} - \hat{p}_{(.)}\right]^2$$

where $\hat{p}_{(.)}$ is the average of the Jackknife estimates:

$$\hat{p}_{(.)} = \frac{\sum_{k=1}^{n} \hat{p}_{(k)}}{n} \quad .$$

Jackknife Variance Estimation

Covariance of Regression Parameters

Start With Given Point Estimator $\hat{\beta}$:
Estimated parameter vector obtained by naively assuming the observations within a cluster are independent

Solution to any score estimating equation of the form

$$\mu(\hat{\beta}) = \sum_{i=1}^{n} \mu_i(\hat{\beta}) = 0$$

where $\mu_i(\hat{\beta})$ is the contribution to the "score" vector from the *i*-th cluster.

Example
Logistic score equations under binomial likelihood

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial \, \mathrm{Log\,L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i \sum w_{ij} \boldsymbol{x}'_{ij} y_{ij} - \sum_i \sum_j w_{ij} \boldsymbol{x}'_{ij} p_{ij}(\boldsymbol{\beta})$$

Jackknife Variance Estimation

Regression Parameters *(continued)*

As long as the model for the marginal mean is correctly specified, the MLE $\hat{\boldsymbol{\beta}}$ is asymptotically consistent and normally distributed

Jackknife Variance Estimator For $\hat{\boldsymbol{\beta}}$

$$Var_{JK}(\hat{\boldsymbol{\beta}}) = \left(\frac{n-p}{n}\right) \sum_{i=1}^{n} \left(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}_{.}\right)\left(\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}_{.}\right)'$$

where

p = number of parameters in the model,

$\hat{\boldsymbol{\beta}}_{-i}$ = estimate of $\boldsymbol{\beta}$ obtained by deleting the m_i observations in PSU i and solving the estimating equations via the Newton-Raphson algorithm, and

$\hat{\boldsymbol{\beta}}_{.}$ = is the average of the $\hat{\boldsymbol{\beta}}_{-i}$.

PSU's are removed sequentially and with-replacement

JK variance estimator is consistent for estimating the asymptotic variance of $\hat{\boldsymbol{\beta}}$

Assumptions and Validity for Taylor Linearization and Jackknife

- PSUs are statistically independent
- No strict distributional assumptions for the response of interest
- Yields consistent estimates of the variance as the number of PSUs tends to infinity
- Method is valid for any underlying intra-PSU correlation structure, as long as PSUs are statistically independent

Balanced Repeated Replication (BRR)

BRR Variance Estimation for Complex Sample Surveys

McCarthy, PJ

1966, *Vital and Health Statistics*, 2(14), NCHS

1969, *Review of the International Statistical Institute* **37**, 239-264.

Wolter, KM

1985, Introduction to Variance Estimation. Springer-Verlag.

Usually assumes PSUs selected ***with-replacement***

Allows for any unbiased sampling method within PSUs

BRR Variance Estimation

How Does It Work?

Balanced Half Samples

- Assume two (2) PSUs are selected with-replacement from each of L strata (more than 2 selections can be made, but is more complicated to explain)

- Form G half-sample replicates, where each half-sample is formed by selecting one of the two PSUs from each stratum based on a Hadamard matrix (Plackett and Burman, 1946)

- Let $\hat{\theta}_g$ be the estimate of the parameter based on the *g-th* half-sample

$$Var(\hat{\theta}) = \frac{1}{G}\sum_{g=1}^{G}(\hat{\theta}_g - \hat{\theta})^2$$

where

$\hat{\theta}$ = estimate based on the full sample

BRR Variance Estimation

BRR Weights: REPWGT *variables*;

- ***Variables*** whose values are the BRR replicate weights for each sampled individual (nonnegative or missing)

- Many survey data bases will supply BRR or replicate weights. SUDAAN assumes the weights are supplied.

Let $w_{gi} = \begin{cases} \textit{replicate weight for sample unit}-i \textit{ in half}-\textit{sample } g \\ 0, \quad \textit{if sample unit}-i \textit{ NOT in half}-\textit{sample } g \end{cases}$

- Use w_{gi} to estimate $\hat{\theta}_g$

For example, the ***Total*** estimate from replicate-*g*:

$$\hat{Y}_g = \sum_{i=1}^{n} w_{gi} y_i$$

- Possible to develop special weights to account for ***without replacement*** sampling. Need to consult with a statistician to develop such weights.

Example Comparing the Three Approaches

These data are taken from a one-year longitudinal study of infant feeding practices of participants in the ***Special Supplemental Nutrition Program for Women, Infants, and Children (WIC)***. A national sample of 42 local agencies (sites) was selected at the first stage and implicitly stratified by region of the country and state within region. Local agencies were paired to form strata. A sample of about 22 pregnant women or new mothers participating in the WIC Program were then selected from each local agency. The participants were interviewed 9 times during each infant's first year of life to gain a complete picture of the feeding patterns of WIC infants. The data consist of one record per WIC respondent.

We use these data to demonstrate the three variance estimation methods in SUDAAN (Taylor linearization, Jackknife, and Balanced Repeated Replication, or BRR). We first estimate descriptive statistics on baby's birth weight and mother's breastfeeding status. Then, we fit a logistic regression model to the incidence of breastfeeding initiation. Point estimates of means, proportions, and regression coefficients are equivalent for all three approaches. Variance estimates are similar in most situations. This example does not point to favoring one method vs. another for variance estimation.

Example Comparing the Three Approaches

Descriptive Statistics

Proportion Initiating Breastfeeding	Sample Size	Percentage	Standard Errors		
			Taylor	Jackknife	BRR
TOTAL	953	54 %	3.1 %	3.1 %	3.1 %
White	480	51 %	3.2 %	3.2 %	3.2 %
African American	225	32 %	4.5 %	4.5 %	4.7 %
Latina	190	83 %	3.5 %	3.5 %	3.5 %
Other	58	63 %	12.2 %	12.8 %	15.0 %

Logistic Regression

Breastfeeding = baby's weight + sex + race + education + marital status

Effect	df	Taylor		Jackknife		BRR	
		Chi-square	P-value	Chi-square	P-value	Chi-square	P-value
Race	3	39.9	0.000	38.7	0.000	36.5	0.000
Education	2	14.2	0.001	14.0	0.001	11.9	0.003
Marital Status	1	36.4	0.000	35.6	0.000	29.9	0.000
Sex	1	0.5	0.463	0.5	0.464	0.5	0.467
Baby's Weight	1	8.3	0.004	8.0	0.005	7.9	0.005

DESCRIPT Programming Statements for Taylor Linearization (DESIGN=WR)

```
                              S U D A A N
            Software for the Statistical Analysis of Correlated Data
            Copyright      Research Triangle Institute      May 1997
                          Beta Test Release 7.5

1   PROC DESCRIPT DATA="WIC" FILETYPE=SAS  DESIGN=WR  DEFT2 MERGEHI;

2   NEST STRATUM SITE;

3   WEIGHT ANALWGT1;

4   VAR BRFDINIT BABYWGT;

5   SUBGROUP RACEMOM EDUC MRTLSTAT BABYSEX;

6   LEVELS   4       3    2        2;

7   SETENV LABWIDTH=28 COLSPCE=1 COLWIDTH=10 LINESIZE=78 DECWIDTH=4 PAGESIZE=60;

8   PRINT NSUM="SAMPLE SIZE" WSUM="POPULATION SIZE" MEAN SEMEAN="S.E."
          DEFFMEAN="DESIGN EFFECT" / STYLE=NCHS NSUMFMT=F6.0 WSUMFMT=F10.0
          DEFFMEANFMT=F6.2 SEMEANFMT=F7.4;

9   TITLE " "  "STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS"  " "
          "TAYLOR LINEARIZATION VARIANCE ESTIMATION"  " ";

Opened SAS data file C:\TERA\EXAMPLES\WIC.SSD for reading.
Number of observations read    :    953    Weighted count :   506279
Denominator degrees of freedom :     21
```

Note the Strata and PSU variables *STRATUM* and *SITE* on the NEST statement, and the analysis weight variable ANALWGT1 on the WEIGHT statement. There are 953 WIC participants on the file, summing to an estimated 506,279 participants in the US population (this is a slight underestimate, since some sites were not included in this example).

DESCRIPT Results Based on Taylor Linearization

```
Date: 07-07-97             Research Triangle Institute            Page  : 1
Time: 14:41:15               The DESCRIPT Procedure               Table : 1

Variance Estimation Method: Taylor Series (WR)

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

TAYLOR LINEARIZATION VARIANCE ESTIMATION

by: Mother Race.
------------------------------------------------------------------------
Variable
   Mother Race                 SAMPLE POPULATION                    DESIGN
                               SIZE   SIZE             Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                          953     506279     0.5370  0.0310   3.67
   White                          480     254567     0.5104  0.0318   1.94
   African American               225     123217     0.3202  0.0445   2.05
   Latina                         190     106306     0.8324  0.0348   1.65
   Other                           58      22189     0.6319  0.1220   3.71
Baby Weight (ozs.)
   Total                          952     505897   116.6232  0.9249   1.90
   White                          480     254567   118.6759  1.0306   1.17
   African American               225     123217   108.6932  1.6144   1.29
   Latina                         189     105924   120.3653  1.2623   1.18
   Other                           58      22189   119.2458  4.6395   2.22
------------------------------------------------------------------------
```

Here we see that breastfeeding initiation and baby's birth weight are both highest among Latina women and lowest among African American women. The standard errors are obtained through Taylor linearization.

DESCRIPT Results Based on Taylor Linearization

```
Date: 07-07-97              Research Triangle Institute             Page  : 2
Time: 14:41:15                The DESCRIPT Procedure                Table : 2

Variance Estimation Method: Taylor Series (WR)

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

TAYLOR LINEARIZATION VARIANCE ESTIMATION

by: Education.
------------------------------------------------------------------------
Variable
   Education                 SAMPLE POPULATION                    DESIGN
                             SIZE   SIZE             Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                        952     505920     0.5367  0.0310   3.67
   < High School                368     212474     0.5187  0.0520   3.98
   High School                  399     211345     0.4991  0.0323   1.66
   > High School                185      82101     0.6804  0.0460   1.80
Baby Weight (ozs.)
   Total                        951     505539   116.6045  0.9291   1.91
   < High School                368     212474   115.2054  1.3358   1.56
   High School                  399     211345   116.5428  1.3042   1.69
   > High School                184      81719   120.4018  2.3207   2.00
------------------------------------------------------------------------
```

Breastfeeding initiation and baby's birth weight among WIC participants is highest among women with more than a high school education.

DESCRIPT Results Based on Taylor Linearization

```
Date: 07-07-97             Research Triangle Institute             Page  : 3
Time: 14:41:15               The DESCRIPT Procedure                Table : 3

Variance Estimation Method: Taylor Series (WR)

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

TAYLOR LINEARIZATION VARIANCE ESTIMATION

by: Marital Status.
------------------------------------------------------------------------
Variable
   Marital Status             SAMPLE POPULATION                   DESIGN
                              SIZE   SIZE             Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                         952     505897     0.5367  0.0310   3.68
   Currently Married             462     246325     0.6393  0.0317   2.02
   Not Currently Married         490     259572     0.4394  0.0330   2.16
Baby Weight (ozs.)
   Total                         952     505897   116.6232  0.9249   1.90
   Currently Married             462     246325   119.0474  1.5696   2.55
   Not Currently Married         490     259572   114.3227  0.9862   1.19
------------------------------------------------------------------------
```

Breastfeeding initiation and baby's birth weight are also higher among those currently married compared to those not currently married.

DESCRIPT Results Based on Taylor Linearization

```
Date: 07-07-97            Research Triangle Institute            Page  : 4
Time: 14:41:15              The DESCRIPT Procedure               Table : 4

Variance Estimation Method: Taylor Series (WR)

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

TAYLOR LINEARIZATION VARIANCE ESTIMATION

by: Baby Sex.
------------------------------------------------------------------------
Variable
   Baby Sex                    SAMPLE POPULATION                  DESIGN
                               SIZE   SIZE            Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                          953     506279     0.5370  0.0310   3.67
   Boy                            495     254670     0.5295  0.0366   2.66
   Girl                           458     251609     0.5446  0.0374   2.58
Baby Weight (ozs.)
   Total                          952     505897   116.6232  0.9249   1.90
   Boy                            494     254288   118.5878  0.8758   0.90
   Girl                           458     251609   114.6377  1.4677   2.30
------------------------------------------------------------------------
```

Breastfeeding initiation is comparable for boy vs. girl babies.

DESCRIPT Results Based on Jackknife Methods

```
PROC DESCRIPT DATA="WIC" FILETYPE=SAS  DESIGN=JACKKNIFE  MERGEHI;

NEST STRATUM SITE;

WEIGHT ANALWGT1;

VAR BRFDINIT BABYWGT;

SUBGROUP RACEMOM EDUC MRTLSTAT BABYSEX;

LEVELS   4       3    2        2 ;

SETENV LABWIDTH=28 COLSPCE=1 COLWIDTH=10 LINESIZE=78 DECWIDTH=4 PAGESIZE=60;

PRINT NSUM="SAMPLE SIZE" WSUM="POPULATION SIZE" MEAN SEMEAN="S.E."
      DEFFMEAN="DESIGN EFFECT" / STYLE=NCHS NSUMFMT=F6.0 WSUMFMT=F10.0
      DEFFMEANFMT=F6.2 SEMEANFMT=F7.4;

TITLE " "  "STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS"
      " "  "JACKKNIFE VARIANCE ESTIMATION"  " ";


Opened SAS data file C:\TERA\EXAMPLES\WIC.SSD for reading.

Number of observations read    :    953    Weighted count :   506279
Denominator degrees of freedom :     21
```

For ***DESIGN=JACKKNIFE***, we keep the NEST and WEIGHT statements as they were for Taylor linearization.

DESCRIPT Results Based on Jackknife Methods

```
Date: 07-07-97             Research Triangle Institute             Page  : 1
Time: 14:41:15               The DESCRIPT Procedure                Table : 1

Variance Estimation Method: Jackknife

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

JACKKNIFE VARIANCE ESTIMATION

by: Mother Race.
------------------------------------------------------------------------
Variable
   Mother Race                 SAMPLE POPULATION                    DESIGN
                               SIZE   SIZE             Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                          953     506279     0.5370  0.0310   3.68
   White                          480     254567     0.5104  0.0318   1.94
   African American               225     123217     0.3202  0.0449   2.07
   Latina                         190     106306     0.8324  0.0352   1.68
   Other                           58      22189     0.6319  0.1279   4.01
Baby Weight (ozs.)
   Total                          952     505897   116.6232  0.9253   1.90
   White                          480     254567   118.6759  1.0338   1.17
   African American               225     123217   108.6932  1.6389   1.33
   Latina                         189     105924   120.3653  1.2730   1.19
   Other                           58      22189   119.2458  4.8725   2.41
------------------------------------------------------------------------
```

Variance estimates for all Jackknife results are similar to Taylor linearization.

DESCRIPT Results Based on Jackknife Methods

```
Date: 07-07-97              Research Triangle Institute             Page  : 2
Time: 14:41:15                The DESCRIPT Procedure                Table : 2

Variance Estimation Method: Jackknife

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

JACKKNIFE VARIANCE ESTIMATION

by: Education.
------------------------------------------------------------------------
Variable
   Education                 SAMPLE POPULATION                    DESIGN
                             SIZE   SIZE             Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                        952     505920     0.5367  0.0310   3.68
   < High School                368     212474     0.5187  0.0523   4.02
   High School                  399     211345     0.4991  0.0323   1.66
   > High School                185      82101     0.6804  0.0461   1.80
Baby Weight (ozs.)
   Total                        951     505539   116.6045  0.9294   1.92
   < High School                368     212474   115.2054  1.3386   1.56
   High School                  399     211345   116.5428  1.3052   1.69
   > High School                184      81719   120.4018  2.3236   1.99
------------------------------------------------------------------------
```

DESCRIPT Results Based on Jackknife Methods

```
Date: 07-07-97            Research Triangle Institute            Page  : 3
Time: 14:41:15              The DESCRIPT Procedure               Table : 3

Variance Estimation Method: Jackknife

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

JACKKNIFE VARIANCE ESTIMATION

by: Marital Status.
------------------------------------------------------------------------
Variable
   Marital Status            SAMPLE POPULATION                    DESIGN
                             SIZE   SIZE            Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                        952     505897     0.5367  0.0311   3.69
   Currently Married            462     246325     0.6393  0.0318   2.02
   Not Currently Married        490     259572     0.4394  0.0330   2.16
Baby Weight (ozs.)
   Total                        952     505897   116.6232  0.9253   1.90
   Currently Married            462     246325   119.0474  1.5704   2.55
   Not Currently Married        490     259572   114.3227  0.9869   1.19
------------------------------------------------------------------------
```

DESCRIPT Results Based on Jackknife Methods

```
Date: 07-07-97            Research Triangle Institute            Page  : 4
Time: 14:41:15              The DESCRIPT Procedure               Table : 4

Variance Estimation Method: Jackknife

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

JACKKNIFE VARIANCE ESTIMATION

by: Baby Sex.
------------------------------------------------------------------------
Variable
   Baby Sex                  SAMPLE POPULATION                    DESIGN
                             SIZE   SIZE             Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                        953     506279     0.5370  0.0310   3.68
   Boy                          495     254670     0.5295  0.0366   2.66
   Girl                         458     251609     0.5446  0.0374   2.58
Baby Weight (ozs.)
   Total                        952     505897   116.6232  0.9253   1.90
   Boy                          494     254288   118.5878  0.8761   0.90
   Girl                         458     251609   114.6377  1.4697   2.30
------------------------------------------------------------------------
```

DESCRIPT Results Based on BRR

```
19  PROC DESCRIPT DATA="WIC" FILETYPE=SAS  DESIGN=BRR  MERGEHI;

20  WEIGHT ANALWGT1;

21  REPWGT RPL001--RPL024;

22  VAR BRFDINIT BABYWGT;

23  SUBGROUP RACEMOM EDUC MRTLSTAT BABYSEX;

24  LEVELS   4       3    2        2 ;

25  SETENV LABWIDTH=28 COLSPCE=1 COLWIDTH=10 LINESIZE=78 DECWIDTH=4 PAGESIZE=60;

26  PRINT NSUM="SAMPLE SIZE" WSUM="POPULATION SIZE" MEAN SEMEAN="S.E."
          DEFFMEAN="DESIGN EFFECT" / STYLE=NCHS NSUMFMT=F6.0 WSUMFMT=F10.0
          DEFFMEANFMT=F6.2 SEMEANFMT=F7.4;

27  TITLE " "  "STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS"  " "
          "BRR VARIANCE ESTIMATION"  " ";


Opened SAS data file C:\TERA\EXAMPLES\WIC.SSD for reading.

Number of observations read    :    953    Weighted count :   506279
Denominator degrees of freedom :     24
```

For ***DESIGN=BRR***, we ***remove the NEST statement*** and include a statement for the known replicate weights (the REPWGT statement). There are 24 replicate weights in this study.

DESCRIPT Results Based on BRR

```
Date: 07-07-97             Research Triangle Institute            Page  : 1
Time: 14:41:15               The DESCRIPT Procedure               Table : 1

Variance Estimation Method: BRR

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

BRR VARIANCE ESTIMATION

by: Mother Race.
------------------------------------------------------------------------
Variable
   Mother Race                 SAMPLE POPULATION                    DESIGN
                               SIZE   SIZE            Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                         953     506279     0.5370  0.0311   3.70
   White                         480     254567     0.5104  0.0321   1.98
   African American              225     123217     0.3202  0.0469   2.27
   Latina                        190     106306     0.8324  0.0351   1.67
   Other                          58      22189     0.6319  0.1496   5.49
Baby Weight (ozs.)
   Total                         952     505897   116.6232  0.9215   1.88
   White                         480     254567   118.6759  1.0957   1.32
   African American              225     123217   108.6932  1.6363   1.32
   Latina                        189     105924   120.3653  1.3577   1.36
   Other                          58      22189   119.2458  5.5170   3.09
------------------------------------------------------------------------
```

Variance estimates based on BRR are similar to Taylor linearization and Jackknife results. Of course, points estimates of the population mean are the same for all three methods.

DESCRIPT Results Based on BRR

```
Date: 07-07-97            Research Triangle Institute            Page  : 2
Time: 14:41:15              The DESCRIPT Procedure               Table : 2

Variance Estimation Method: BRR

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

BRR VARIANCE ESTIMATION

by: Education.
------------------------------------------------------------------------
Variable
   Education                 SAMPLE POPULATION                    DESIGN
                             SIZE   SIZE            Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                        952     505920     0.5367  0.0311   3.70
   < High School                368     212474     0.5187  0.0521   3.99
   High School                  399     211345     0.4991  0.0337   1.80
   > High School                185      82101     0.6804  0.0466   1.84
Baby Weight (ozs.)
   Total                        951     505539   116.6045  0.9255   1.90
   < High School                368     212474   115.2054  1.3756   1.65
   High School                  399     211345   116.5428  1.3217   1.73
   > High School                184      81719   120.4018  2.4006   2.13
------------------------------------------------------------------------
```

DESCRIPT Results Based on BRR

```
Date: 07-07-97            Research Triangle Institute            Page  : 3
Time: 14:41:15              The DESCRIPT Procedure               Table : 3

Variance Estimation Method: BRR

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

BRR VARIANCE ESTIMATION

by: Marital Status.
------------------------------------------------------------------------
Variable
   Marital Status            SAMPLE POPULATION                    DESIGN
                             SIZE   SIZE            Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                        952     505897     0.5367  0.0311   3.71
   Currently Married            462     246325     0.6393  0.0316   2.00
   Not Currently Married        490     259572     0.4394  0.0338   2.27
Baby Weight (ozs.)
   Total                        952     505897   116.6232  0.9215   1.88
   Currently Married            462     246325   119.0474  1.6293   2.75
   Not Currently Married        490     259572   114.3227  0.9802   1.17
------------------------------------------------------------------------
```

DESCRIPT Results Based on BRR

```
Date: 07-07-97            Research Triangle Institute             Page  : 4
Time: 14:41:15              The DESCRIPT Procedure                Table : 4

Variance Estimation Method: BRR

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

BRR VARIANCE ESTIMATION

by: Baby Sex.
------------------------------------------------------------------------
Variable
   Baby Sex                    SAMPLE POPULATION                   DESIGN
                               SIZE   SIZE            Mean    S.E. EFFECT
------------------------------------------------------------------------
Breastfeeding Initiation
   Total                          953     506279    0.5370  0.0311   3.70
   Boy                            495     254670    0.5295  0.0374   2.77
   Girl                           458     251609    0.5446  0.0367   2.49
Baby Weight (ozs.)
   Total                          952     505897  116.6232  0.9215   1.88
   Boy                            494     254288  118.5878  0.9277   1.01
   Girl                           458     251609  114.6377  1.3999   2.09
------------------------------------------------------------------------
```

LOGISTIC Modelling Based on Taylor Linearization

```
28  PROC LOGISTIC DATA="WIC" FILETYPE=SAS  DESIGN=WR  DEFT2 MERGEHI;

29  NEST STRATUM SITE;

30  WEIGHT ANALWGT1;

31  SUBGROUP RACEMOM EDUC MRTLSTAT BABYSEX;

32  LEVELS   4       3    2        2   ;

33  REFLEVEL RACEMOM=1 EDUC=1;

34  MODEL BRFDINIT = BABYWGT BABYSEX RACEMOM EDUC MRTLSTAT;

35  EFFECTS RACEMOM=(0 1 -1 0) / NAME="African Am Vs. Latina";

36  TEST WALDCHI;

37  SETENV COLSPCE=2 LABWIDTH=26 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

38  PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" OR LOWOR UPOR
          DF="DF" WALDCHI="WALD CHI-SQ" WALDCHP="P-VALUE"
          / T_BETAFMT=F8.2 DEFTFMT=F6.2 SEBETAFMT=F8.6
           ORFMT=F5.2 LOWORFMT=F6.2 UPORFMT=F6.2
           DFFMT=F7.0 WALDCHIFMT=F8.2 ;

39  TITLE " "  "STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS"  " "
          "TAYLOR SERIES VARIANCE ESTIMATION";

NOTE: Terms in the MODEL statement have been rearranged
      to follow subgroup order.

Opened SAS data file C:\TERA\EXAMPLES\WIC.SSD for reading.

Number of observations read       :    953    Weighted count:   506279
Observations used in the analysis :    951    Weighted count:   505539
Observations with missing values  :      2    Weighted count:      740
Denominator degrees of freedom    :     21

Maximum number of estimable parameters for the model is  9
Number of zero responses     :   431
Number of non-zero responses :   520

Parameters have converged in 4 iterations

R-Square for dependent variable BRFDINIT (Cox & Snell, 1989): 0.171348
```

In the logistic models, we want to see if baby's birth weight, sex, as well as the mother's race, education, and marital status significantly affect breastfeeding initiation. More than half the sample initiated breastfeeding (520 out of 951 non-missing responses). The SUBGROUP and

LEVELS statements define the variables to be treated as categorical, and the REFLEVEL statement changes the default reference levels for two of the categorical covariates from the last level to the first level. The EFFECTS statement directly compares African American women to Latina women.

LOGISTIC Modelling Based on Taylor Linearization

```
Date: 07-07-97            Research Triangle Institute            Page  : 1
Time: 14:41:15              The LOGISTIC Procedure               Table : 1

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

TAYLOR SERIES VARIANCE ESTIMATION

---------------------------------------------------------------------------
Independent Variables and
  Effects                                        DESIGN
                                BETA      S.E.   EFFECT  T:BETA=0   P-VALUE
---------------------------------------------------------------------------
Intercept                    -1.5928  0.427856    0.96     -3.72    0.0013
Mother Race
  White                       0.0000  0.000000     .         .        .
  African American           -0.5410  0.233989    1.70     -2.31    0.0310
  Latina                      1.7147  0.300657    1.80      5.70    0.0000
  Other                       0.4155  0.547545    2.26      0.76    0.4563
Education
  < High School               0.0000  0.000000     .         .        .
  High School                 0.0565  0.210793    1.81      0.27    0.7911
  > High School               0.8533  0.298020    1.75      2.86    0.0093
Marital Status
  Currently Married           0.6843  0.113357    0.57      6.04    0.0000
  Not Currently Married       0.0000  0.000000     .         .        .
Baby Sex
  Boy                        -0.1232  0.167831    1.33     -0.73    0.4709
  Girl                        0.0000  0.000000     .         .        .
Baby Weight (ozs.)            0.0096  0.003325    0.89      2.89    0.0089
---------------------------------------------------------------------------
```

From the estimated regression coefficients we see immediately that significantly fewer African American women, but significantly more Latina women, initiated breastfeeding compared to white women. Also, having more than a high school education and being currently married both significantly improved the likelihood of breastfeeding. Finally, as baby's birth weight increased, the likelihood of breastfeeding was significantly increased.

LOGISTIC Modelling Based on Taylor Linearization

```
Date: 07-07-97            Research Triangle Institute            Page  : 2
Time: 14:41:15              The LOGISTIC Procedure               Table : 1

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

TAYLOR SERIES VARIANCE ESTIMATION

-------------------------------------------------------
Contrast                             WALD
                                 DF  CHI-SQ     P-VALUE
-------------------------------------------------------
OVERALL MODEL                     9    144.66    0.0000
MODEL MINUS INTERCEPT             8    135.98    0.0000
INTERCEPT                         .       .        .
RACEMOM                           3     39.92    0.0000
EDUC                              2     14.15    0.0008
MRTLSTAT                          1     36.44    0.0000
BABYSEX                           1      0.54    0.4628
BABYWGT                           1      8.33    0.0039
African Am Vs. Latina             1     37.56    0.0000
-------------------------------------------------------
```

Under Taylor linearization, mother’s race, education, marital status, and baby’s birth weight were all statistically significant. Also, the user-specified contrast comparing African American women to Latina women was statistically significant.

LOGISTIC Modelling Based on Taylor Linearization

```
Date: 07-07-97            Research Triangle Institute            Page  : 3
Time: 14:41:15              The LOGISTIC Procedure               Table : 1

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

TAYLOR SERIES VARIANCE ESTIMATION

-------------------------------------------------
Independent Variables and
  Effects                          Lower   Upper
                            Odds   95%     95%
                            Ratio  Limit   Limit
-------------------------------------------------
Intercept                    0.20    0.08    0.50
Mother Race
  White                      1.00    1.00    1.00
  African American           0.58    0.36    0.95
  Latina                     5.55    2.97   10.38
  Other                      1.52    0.49    4.73
Education
  < High School              1.00    1.00    1.00
  High School                1.06    0.68    1.64
  > High School              2.35    1.26    4.36
Marital Status
  Currently Married          1.98    1.57    2.51
  Not Currently Married      1.00    1.00    1.00
Baby Sex
  Boy                        0.88    0.62    1.25
  Girl                       1.00    1.00    1.00
Baby Weight (ozs.)           1.01    1.00    1.02
-------------------------------------------------

LOGISTIC used
  CPU time       : 4.83 seconds
  Elapsed time   : 5 seconds
  Virtual memory : 2.40 MB
```

The estimated odds ratios and 95% confidence limits indicate that, for example:

- the odds of initiated breastfeeding are increased by more than five-fold for Latina women vs. white women
- the odds are reduced by half in African American women vs. white women
- the odds are approximately doubled for women who are currently married as well as for women with more than a high school education.

LOGISTIC Modelling Based on Jackknife Methods

```
40  PROC LOGISTIC DATA="WIC" FILETYPE=SAS  DESIGN=JACKKNIFE  MERGEHI;

41  NEST STRATUM SITE;

42  WEIGHT ANALWGT1;

43  SUBGROUP RACEMOM EDUC MRTLSTAT BABYSEX;

44  LEVELS   4       3    2        2  ;

45  REFLEVEL RACEMOM=1 EDUC=1;

46  MODEL BRFDINIT = BABYWGT BABYSEX RACEMOM EDUC MRTLSTAT;

47  EFFECTS RACEMOM=(0 1 -1 0) / NAME="African Am Vs. Latina";

48  TEST WALDCHI;

49  SETENV COLSPCE=2 LABWIDTH=26 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

50  PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" OR LOWOR UPOR
          DF="DF" WALDCHI="WALD CHI-SQ" WALDCHP="P-VALUE"
          / T_BETAFMT=F8.2 DEFTFMT=F6.2 SEBETAFMT=F8.6
           ORFMT=F5.2 LOWORFMT=F6.2 UPORFMT=F6.2
           DFFMT=F7.0 WALDCHIFMT=F8.2 ;

51  TITLE " "  "STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS"  " "
          "JACKKNIFE VARIANCE ESTIMATION";

NOTE: Terms in the MODEL statement have been rearranged
      to follow subgroup order.

Opened SAS data file C:\TERA\EXAMPLES\WIC.SSD for reading.

Number of observations read       :    953    Weighted count:   506279
Observations used in the analysis :    951    Weighted count:   505539
Observations with missing values  :      2    Weighted count:      740
Denominator degrees of freedom    :     21

Maximum number of estimable parameters for the model is  9
Number of zero responses     :   431
Number of non-zero responses :   520

Parameters have converged in 4 iterations

R-Square for dependent variable BRFDINIT (Cox & Snell, 1989): 0.171348
```

The results of logistic modelling using the Jackknife variance estimation method are very similar to Taylor linearization.

LOGISTIC Modelling Based on Jackknife Methods

```
Date: 07-07-97             Research Triangle Institute              Page  : 1
Time: 14:41:15               The LOGISTIC Procedure                 Table : 1
```

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

JACKKNIFE VARIANCE ESTIMATION

Independent Variables and Effects	BETA	S.E.	DESIGN EFFECT	T:BETA=0	P-VALUE
Intercept	-1.5928	0.432309	0.94	-3.68	0.0014
Mother Race					
White	0.0000	0.000000	.	.	.
African American	-0.5410	0.235709	1.69	-2.30	0.0321
Latina	1.7147	0.306057	1.96	5.60	0.0000
Other	0.4155	0.601481	3.00	0.69	0.4972
Education					
< High School	0.0000	0.000000	.	.	.
High School	0.0565	0.215382	1.83	0.26	0.7955
> High School	0.8533	0.302420	1.92	2.82	0.0102
Marital Status					
Currently Married	0.6843	0.114777	0.59	5.96	0.0000
Not Currently Married	0.0000	0.000000	.	.	.
Baby Sex					
Boy	-0.1232	0.168455	1.35	-0.73	0.4725
Girl	0.0000	0.000000	.	.	.
Baby Weight (ozs.)	0.0096	0.003393	0.88	2.83	0.0101

LOGISTIC Modelling Based on Jackknife Methods

```
Date: 07-07-97            Research Triangle Institute            Page  : 2
Time: 14:41:15              The LOGISTIC Procedure               Table : 1

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

JACKKNIFE VARIANCE ESTIMATION

-------------------------------------------------------
Contrast                             WALD
                                DF  CHI-SQ     P-VALUE
-------------------------------------------------------
OVERALL MODEL                    9    142.32    0.0000
MODEL MINUS INTERCEPT            8    134.01    0.0000
INTERCEPT                        .       .         .
RACEMOM                          3     38.67    0.0000
EDUC                             2     14.01    0.0009
MRTLSTAT                         1     35.55    0.0000
BABYSEX                          1      0.54    0.4644
BABYWGT                          1      7.99    0.0047
African Am Vs. Latina            1     36.45    0.0000
-------------------------------------------------------
```

LOGISTIC Modelling Based on Jackknife Methods

```
Date: 07-07-97            Research Triangle Institute            Page  : 3
Time: 14:41:15              The LOGISTIC Procedure               Table : 1
```

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

JACKKNIFE VARIANCE ESTIMATION

Independent Variables and Effects	Odds Ratio	Lower 95% Limit	Upper 95% Limit
Intercept	0.20	0.08	0.50
Mother Race			
White	1.00	1.00	1.00
African American	0.58	0.36	0.95
Latina	5.55	2.94	10.50
Other	1.52	0.43	5.29
Education			
< High School	1.00	1.00	1.00
High School	1.06	0.68	1.66
> High School	2.35	1.25	4.40
Marital Status			
Currently Married	1.98	1.56	2.52
Not Currently Married	1.00	1.00	1.00
Baby Sex			
Boy	0.88	0.62	1.25
Girl	1.00	1.00	1.00
Baby Weight (ozs.)	1.01	1.00	1.02

```
LOGISTIC used
  CPU time       : 9.12 seconds
  Elapsed time   : 10 seconds
  Virtual memory : 2.33 MB
```

LOGISTIC Modelling Based on BRR

```
52  PROC LOGISTIC DATA="WIC" FILETYPE=SAS  DESIGN=BRR  MERGEHI;

53  WEIGHT ANALWGT1;

54  REPWGT RPL001--RPL024;

55  SUBGROUP RACEMOM EDUC MRTLSTAT BABYSEX ;

56  LEVELS   4       3    2        2   ;

57  REFLEVEL RACEMOM=1 EDUC=1;

58  MODEL BRFDINIT = BABYWGT BABYSEX RACEMOM EDUC MRTLSTAT;

59  EFFECTS RACEMOM=(0 1 -1 0) / NAME="African Am Vs. Latina";

60  TEST WALDCHI;

61  SETENV COLSPCE=2 LABWIDTH=26 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

62  PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" OR LOWOR UPOR
          DF="DF" WALDCHI="WALD CHI-SQ" WALDCHP="P-VALUE"
          / T_BETAFMT=F8.2 DEFTFMT=F6.2 SEBETAFMT=F8.6
           ORFMT=F5.2 LOWORFMT=F6.2 UPORFMT=F6.2
           DFFMT=F7.0 WALDCHIFMT=F8.2 ;

63  TITLE " "  "STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS"  " "
          "BRR VARIANCE ESTIMATION";

NOTE: Terms in the MODEL statement have been rearranged
      to follow subgroup order.

Opened SAS data file C:\TERA\EXAMPLES\WIC.SSD for reading.

Number of observations read       :    953    Weighted count:   506279
Observations used in the analysis :    951    Weighted count:   505539
Observations with missing values  :      2    Weighted count:      740
Denominator degrees of freedom    :     24

Maximum number of estimable parameters for the model is  9
Number of zero responses     :   431
Number of non-zero responses :   520

Parameters have converged in 4 iterations

R-Square for dependent variable BRFDINIT (Cox & Snell, 1989): 0.171348
```

Logistic modelling results using BRR variance estimation methods are similar to those based on Taylor linearization and the Jackknife.

LOGISTIC Modelling Based on BRR

```
Date: 07-07-97            Research Triangle Institute            Page  : 1
Time: 14:41:15              The LOGISTIC Procedure               Table : 1

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

BRR VARIANCE ESTIMATION

--------------------------------------------------------------------------
Independent Variables and
  Effects                                        DESIGN
                                BETA      S.E.   EFFECT  T:BETA=0   P-VALUE
--------------------------------------------------------------------------
Intercept                    -1.5928  0.442342    0.98     -3.60    0.0014
Mother Race
  White                       0.0000  0.000000     .         .        .
  African American           -0.5410  0.232914    1.65     -2.32    0.0290
  Latina                      1.7147  0.307665    1.98      5.57    0.0000
  Other                       0.4155  1.277375   13.53      0.33    0.7478
Education
  < High School               0.0000  0.000000     .         .        .
  High School                 0.0565  0.219845    1.90      0.26    0.7992
  > High School               0.8533  0.311799    2.04      2.74    0.0115
Marital Status
  Currently Married           0.6843  0.125232    0.71      5.46    0.0000
  Not Currently Married       0.0000  0.000000     .         .        .
Baby Sex
  Boy                        -0.1232  0.169507    1.37     -0.73    0.4742
  Girl                        0.0000  0.000000     .         .        .
Baby Weight (ozs.)            0.0096  0.003416    0.89      2.81    0.0097
--------------------------------------------------------------------------
```

LOGISTIC Modelling Based on BRR

```
Date: 07-07-97            Research Triangle Institute             Page  : 2
Time: 14:41:15              The LOGISTIC Procedure                Table : 1
```

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

BRR VARIANCE ESTIMATION

Contrast	DF	WALD CHI-SQ	P-VALUE
OVERALL MODEL	9	111.03	0.0000
MODEL MINUS INTERCEPT	8	108.99	0.0000
INTERCEPT	.	.	.
RACEMOM	3	36.48	0.0000
EDUC	2	11.92	0.0026
MRTLSTAT	1	29.86	0.0000
BABYSEX	1	0.53	0.4672
BABYWGT	1	7.89	0.0050
African Am Vs. Latina	1	33.99	0.0000

LOGISTIC Modelling Based on BRR

```
Date: 07-07-97            Research Triangle Institute            Page  : 3
Time: 14:41:15              The LOGISTIC Procedure               Table : 1

Response variable BRFDINIT: Breastfeeding Initiation

STUDY OF BREAST-FEEDING PATTERNS AMONG WIC PARTICIPANTS

BRR VARIANCE ESTIMATION

-------------------------------------------------
Independent Variables and
  Effects                          Lower   Upper
                            Odds   95%     95%
                            Ratio  Limit   Limit
-------------------------------------------------
Intercept                    0.20    0.08    0.51
Mother Race
  White                      1.00    1.00    1.00
  African American           0.58    0.36    0.94
  Latina                     5.55    2.94   10.48
  Other                      1.52    0.11   21.15
Education
  < High School              1.00    1.00    1.00
  High School                1.06    0.67    1.67
  > High School              2.35    1.23    4.47
Marital Status
  Currently Married          1.98    1.53    2.57
  Not Currently Married      1.00    1.00    1.00
Baby Sex
  Boy                        0.88    0.62    1.25
  Girl                       1.00    1.00    1.00
Baby Weight (ozs.)           1.01    1.00    1.02
-------------------------------------------------

LOGISTIC used
  CPU time       : 6.43 seconds
  Elapsed time   : 7 seconds
  Virtual memory : 2.41 MB
```

The MULTILOG Procedure

Multinomial Logistic Regression
(Release 7.0)

- ***Generalized Logit Models***

  - Nominal Outcomes

    e.g., Type of health plan (A, B, C, D)

- ***Cumulative Logit Models***

  - Ordinal Outcomes

    e.g., Pain Relief:
    none, mild, moderate, complete relief

  - "Proportional Odds Models"

- Binary Logistic is a special case of each

- Model-fitting Approach

  - Fits ***marginal*** or ***population-averaged*** models

  - Uses GEE to model the intracluster correlations and efficiently estimate regression coefficients

Applications in Pharmaceutical Research

Toxicology / Pre-Clinical Studies

- ***Developmental Toxicity***
  Severity of malformations recorded on fetuses clustered within litters (cluster = litter)

Clinical Trials

- ***Repeated Measures Studies***
  Multiple illness or adverse events per patient
  (cluster = patient)

  Example
  Repeated ordinal responses of pain relief over an 8-hour period in a randomized clinical trial of acute pain relief comparing placebo with 2 analgesics (Gansky, Koch, et al., 1994, Journal of Biopharmaceutical Statistics)

- ***Cross-Over Studies***
  Subjects receive each treatment in sequence
  (cluster = patient)

  Example
  3-period, 3 treatment cross-over study (Snapinn and Small, 1986, Biometrics):
  Investigational drug, aspirin, and placebo administered in sequence to headache sufferers
  Patients rated each drug on scale of 1-4 according to amount of pain relief.

Generalized Logit Model

Y is a categorical response variable with *K* categories 1,2,...,*K* (nominal scale)

$\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})' =$ vector of explanatory variables for subject *i*

Model $\pi_k(\boldsymbol{x_i}) = prob(Y_i = k \mid \boldsymbol{x_i}) \qquad k = 1, \ldots, K-1$

Generalized Logits Model **(Agresti, 1990):**

$$\log\left[\frac{\pi_k(\boldsymbol{x}_i)}{\pi_K(\boldsymbol{x}_i)}\right] = \boldsymbol{\beta}'_k \boldsymbol{x}_i \qquad k = 1, \ldots, K-1$$

- Separate parameter vector (intercepts and slopes) for *each* of the *K-1* logit equations
- $\boldsymbol{\beta}_K = \mathbf{0}$.
- $\exp(\boldsymbol{\beta}_k)$ = odds of being in category *k* vs. *K* (the last) for each 1-unit increase in *x*

Cumulative Logit Model

Y is a categorical response variable with K categories 1,2,...,K

ordinal scale: *e.g.*, none, mild, moderate, severe

$\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})' =$ vector of explanatory variables for subject i

Model $F_k(\boldsymbol{x_i}) = prob(Y \le k \mid \boldsymbol{x_i}) =$ cum. prob. up to and including category k

McCullagh's (1980) Proportional Odds Model:

Cumulative Logits

$$\log\left[\frac{F_k(\boldsymbol{x}_i)}{1 - F_k(\boldsymbol{x}_i)}\right] = \alpha_k + \boldsymbol{\beta}'\boldsymbol{x}_i \qquad k = 1, \ldots, K-1$$

- Separate intercepts $\boldsymbol{\alpha_k}$, but a ***common set of slopes*** $\boldsymbol{\beta}$, for k = 1,...,K-1
- $\boldsymbol{\beta}$ measures the effect of the covariates on the severity of response

Efficient Parameter Estimation

Efficiently Weight the Data to Estimate Regression Coefficients (β)

GEE Approach

(Longitudinal Data Analysis, Zeger and Liang, 1986):

1) Assume a Covariance Structure $\boldsymbol{V_i}$ to describe the relationship among observations within clusters, i=1,...,n

   - Mean / Variance Relationship:
     $V(y_{ij}) = g(\mu_{ij})$

   - Pairwise Correlation Model:
     $Corr(y_{ij}, y_{ik})$

2) Estimate Covariance Parameters

3) Weight Data Inversely Proportional to V_i to Estimate β

   V_i inserted into the usual estimating equations in order to weight the data efficiently

Efficient Parameter Estimation

Efficiently Weight the Data to Estimate Regression Coefficients (β)

GEE Approach

(Longitudinal Data Analysis, Zeger and Liang, 1986):

$i = 1 , \ldots , n$	Clusters
$j = 1 , \ldots , m_i$	Observational Units
$\boldsymbol{y_i} = (y_{i1}, \ldots, y_{im_i})$	Vector of responses
$\boldsymbol{\mu_i} = E(\boldsymbol{y_i}) = \boldsymbol{\mu_i}(\boldsymbol{\beta}) = (\mu_{i1}, \ldots, \mu_{im_i})$	Vector of marginal means
$\boldsymbol{V_i}(\boldsymbol{\alpha}) = Cov(\boldsymbol{y_i} ; \boldsymbol{\mu_i} , \alpha)$	Working Covariance matrix

"Generalized" Estimating Equations:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \boldsymbol{V_i}(\boldsymbol{\alpha})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

Working Covariance Structure

$$\boldsymbol{V}_i(\alpha) = \boldsymbol{A}_i^{1/2}\, \boldsymbol{R}_i(\boldsymbol{\alpha})\, \boldsymbol{A}_i^{1/2} \cdot \phi \qquad \boldsymbol{V} \text{ is Block diagonal}$$

$\boldsymbol{A}_i$ = diagonal matrix with diagonal elements equal to the marginal variances of observational units within clusters: $g(\mu_{i1}), \ldots, g(\mu_{i m_i})$

$$= \begin{bmatrix} g(\mu_{i1}) & 0 & 0 & 0 \\ 0 & g(\mu_{i2}) & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & g(\mu_{i m_i}) \end{bmatrix}$$

Relationship Between Variance of y_{ij} and its Mean

$$Var(y_{ij}) = g(\mu_{ij}) \cdot \phi$$

g is a known variance function, ϕ is an unknown scale parameter

Binary Responses

Marginal distribution of y_{ij} is Bernoulli

Therefore $Var(y_{ij}) = \mu_{ij}(1 - \mu_{ij})$ and $\phi = 1$.

Choices for Working Correlation Matrices

$R_i(\alpha)$ is the "Working" Correlation Matrix for y_i

$$\alpha_{jk} = corr\,(y_{ij}, y_{ik})$$

1) ***Independent Working Correlation Matrix***
(Identity matrix implies 0 pairwise correlation)

$$R_i(\alpha) = I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Estimating equations reduce to familiar forms:

  - Normal equations for linear regression
  - Score equations for logistic regression

- Leads to standard regression coefficient estimates

- Consistent and asymptotically normal, regardless of whether or not the correlation structure is correctly specified

- This approach is offered in SUDAAN, and it is perfectly valid for estimating the ***regression parameters*.**

Choices for Working Correlation Matrices

2) ***Exchangeable***
(equal pairwise correlations)

$$\boldsymbol{R_i}(\boldsymbol{\alpha}) = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

- SUDAAN offers this form as well

- Can improve ***efficiency*** of parameter estimates over the independence working assumption when working correlations are close to truth.

Robust Variance Estimate for GEE

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = M_0^{-1} \, M_1 \, M_0^{-1}$$

where

$$M_0 = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \, \boldsymbol{V}_i^{-1} \, \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$

$$M_1 = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \, \boldsymbol{V}_i^{-1} \, Var(\boldsymbol{y}_i) \, \boldsymbol{V}_i^{-1} \, \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$

- M_0^{-1} (outside term) is called the ***naive*** or ***model-based*** variance (inverse of information matrix, appropriate when working assumption about covariance structure is correct)

  Sensitive to violations of model assumptions!

- M_1 (middle term) serves as a ***variance correction*** when the covariance model is misspecified

- ***Robust variance*** is consistent even when $var(y_{ij}) \neq g(\mu_{ij}) \cdot \phi$ or $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is not the true correlation matrix of $\boldsymbol{Y}_i$

- $Var(\boldsymbol{y}_i)$ empirically estimated by $(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)'$

- SUDAAN offers the ***robust*** (default) and in Release 7.5 the ***model-based*** variance estimates (via the ***SEMETHOD=MODEL*** option)

Robust Variance Estimate for GEE

- Also referred to as ***Sandwich Estimator*** or ***Variance Correction***

- Properly accounts for intracluster correlation

- Yields ***consistent variance estimates***, even if correlation structure is misspecified (*e.g.*, by specifying "working" independence when the correlations are in fact exchangeable)

  Huber (1967)
  Royall (1986)
  Binder (1983, 1992)

SYNTAX for GEE options in REGRESS and MULTILOG

```
PROC REGRESS
     MULTILOG ... R = Independent | Exchangeable
                  RSTEPS = count
                  SEMETHOD = ZEGER | BINDER | MODEL
```

R = ***Independent* | *Exchangeable***

Specifies the "working" assumption for estimating the within-cluster correlation structure. The default assumption is independent working correlations. When *R=exchangeable*, the estimated exchangeable correlation matrix is available for printing.

RSTEPS = ***count***

Specifies the maximum number of steps (iterating between estimated regression coefficients and correlations) used to fit the model. The default value is 0 and the default correlation structure is independent (*R=independent*). If you specify exchangeable correlations, the default value for the RSTEPS parameter is 1.

SEMETHOD = ***ZEGER* | *BINDER* | *MODEL***

Specifies the method for computing standard errors of regression coefficients. ***SEMETHOD=ZEGER*** and ***BINDER*** both specify the full ***robust*** or ***sandwich*** variance estimator. For the REGRESS procedure, *ZEGER* and *BINDER* produce identical results. For the MULTILOG procedure, *ZEGER* and *BINDER* produce different results for responses with more than 2 levels. ***SEMETHOD=MODEL*** requests the ***model-based*** or ***naive*** standard error estimator, which is simply the outside of the sandwich estimator and is appropriate when the pairwise correlations within a cluster have been correctly specified.

What Does SUDAAN Model?

Marginal Models (Population-Averaged)

- *Marginal mean* of the multivariate outcomes as a function of the covariates:

$$F\left[E(y_{ij} \mid \boldsymbol{x_{ij}})\right] = \boldsymbol{x}_{ij}^{\prime}\boldsymbol{\beta}$$

- Focus on how X causes Y, while acknowledging the dependence within clusters (as opposed to how one Y causes another)

- Describes relationship between covariates and response *across* clusters

- Intracluster correlation treated as nuisance parameter

References:

Zeger and Liang (1986)
Liang and Zeger (1986)
Zeger, Liang, and Albert (1988)
Binder (1983, 1992)

R-Square for Logistic Regression

Proportion of Log-Likelihood Explained by the Model
(Cox and Snell, 1989)

$$R^2 = 1 - \left(\frac{L(\mathbf{0})}{L(\hat{\boldsymbol{\beta}})} \right)^{\frac{2}{n}}$$

where:

$L(\mathbf{0})$ is the likelihood of the intercept-only model
$L(\hat{\boldsymbol{\beta}})$ is the likelihood of the specified model, and
n is the sample size.

R-Square for *Linear* Regression:
Simple correlation between observed and predicted response
(based on the model).

REFLEVEL Statement

- Available in all modelling procedures

- Allows the user to change the definition of the ***reference cell*** for all categorical covariates.

- By ***default***, the reference cell is the ***last level*** of each categorical covariate.

Syntax:

```
REFLEVEL  variable_1 = reference_level_1
          variable_2 = reference_level_2

          {... variable_k = reference_level_k};
```

- Each *variable_i* must be defined on the SUBGROUP and LEVELS statements

- For SUBGROUP variables ***not*** on the REFLEVEL statement, the default reference level is still the ***last*** level.

REFLEVEL Example

The following example comes from the NHANES I Survey and its Longitudinal Follow-up Study conducted 10 years later. NHANES I (*National Health and Nutrition Examination Survey I*) was a multi-stage sample survey of over 14,000 adults in the US aged 25-74 years, with data collection taking place in 1971-1975. The epidemiologic follow-up took place in 1981-1984.

In this analysis, we wish to determine whether follow-up cancer status (***CANCER12***, 1=*yes* vs. 0=*no*) is associated with a measure of body iron stores at the initial exam (***B_TIBC***, total iron-binding capacity), while adjusting for age group at initial exam (***AGEGROUP***, 1=*20-49*, 2=*50+*) and smoking status (***SMOKE***, 1=*current*, 2=*former*, 3=*never*, 4=*unknown*).

First, we supply the results with the ***default reference cells***, the last level of each categorical covariate, *i.e.*, SMOKE=4 (*unknown*) and AGEGROUP=2 (*50+)*:

```
1   PROC MULTILOG DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

2   NEST Q_STRATA PSU1;

3   WEIGHT B_WTIRON;

4   SUBGROUP CANCER12 AGEGROUP SMOKE;

5   LEVELS   2        2        4;

6   MODEL CANCER12 = B_TIBC AGEGROUP SMOKE / CUMLOGIT;

7   SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

8   PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" DF WALDCHI WALDCHP / T_BETAFMT=F8.2 DEFTFMT=F6.2
          WALDCHIFMT=F8.2 DFFMT=F8.0;

9   TITLE "Default Reference Cell Model";

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.

Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35
Maximum number of estimable parameters for the model is  6

File C:\ADVANCED\IRONSUD.SSD contains   67 Clusters
Maximum cluster size is 111 records
Minimum cluster size is  15 records
Independence parameters have converged in 5 iterations

Sample and Population Counts for Response Variable CANCER12
  Cancer   :  Sample Count      232    Population Count   1745695
  No Cancer:  Sample Count     3058    Population Count  38824628
```

REFLEVEL Example

DEFAULT Reference Cell Parameterization

```
Date: 05-29-97            Research Triangle Institute            Page  : 1
Time: 14:16:21              The MULTILOG Procedure               Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable CANCER12: Cancer Status (1/2)

Default Reference Cell Model

----------------------------------------------------------------------
Independent Variables                               DESIGN
  and Effects                         BETA     S.E. EFFECT T:BETA=0  P-VALUE
----------------------------------------------------------------------
Intercept                          -0.8618   0.6605   0.94    -1.30   0.2004
Total Iron Binding Capacity        -0.0024   0.0018   1.10    -1.29   0.2052
Age Cohort
  20-49 yrs.                       -2.2525   0.3343   1.89    -6.74   0.0000
  50+ yrs.                          0.0000   0.0000    .        .       .
Smoking Status
  Current                          -0.5858   0.2771   0.77    -2.11   0.0417
  Former                           -0.9418   0.2922   0.84    -3.22   0.0027
  Never                            -0.4998   0.2743   0.85    -1.82   0.0770
  Unknown                           0.0000   0.0000    .        .       .
----------------------------------------------------------------------
```

Here, each smoking group is automatically compared to the ***unknown*** smoking status (SMOKE=4), which may not be very meaningful.

REFLEVEL Example

DEFAULT Reference Cell Parameterization

```
Date: 05-29-97             Research Triangle Institute            Page  : 2
Time: 14:16:21               The MULTILOG Procedure               Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable CANCER12: Cancer Status (1/2)

Default Reference Cell Model

----------------------------------------------------
Contrast                  Degrees           P-value
                          of       Wald     Wald
                          Freedom  ChiSq    ChiSq
----------------------------------------------------
OVERALL MODEL                   6   708.28   0.0000
MODEL MINUS INTERCEPT           5    64.47   0.0000
B_TIBC                          1     1.67   0.1967
AGEGROUP                        1    45.39   0.0000
SMOKE                           3    10.60   0.0141
----------------------------------------------------

MULTILOG used
  CPU time       : 12.74 seconds
  Elapsed time   : 13 seconds
  Virtual memory : 2.84 MB
```

Here we see that ***Age group*** and ***Smoking status*** are significantly associated with follow-up cancer status, but ***Total iron-binding capacity*** is not (*p*=0.1967).

REFLEVEL Example

Using the REFLEVEL Statement

Next, using the REFLEVEL statement, we re-define the reference cells to be the ***first level*** of each categorical variable. Note the only differences in the results are in the estimates of the regression coefficients, where the expected value of the response for each level of the categorical covariate(s) is now compared to the user-specified *first* level instead of the last. The main effects tests remain unchanged.

```
10  PROC MULTILOG DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

11  NEST Q_STRATA PSU1;

12  WEIGHT B_WTIRON;

13  REFLEVEL AGEGROUP=1 SMOKE=1;

14  SUBGROUP CANCER12 AGEGROUP SMOKE;

15  LEVELS   2        2        4;

16  MODEL CANCER12 = B_TIBC AGEGROUP SMOKE / CUMLOGIT;

17  SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

18  PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" DF WALDCHI WALDCHP / T_BETAFMT=F8.2 DEFTFMT=F6.2
          WALDCHIFMT=F8.2 DFFMT=F8.0;

19  TITLE "Using the REFLEVEL Statement";

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.

Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is  6

File C:\ADVANCED\IRONSUD.SSD contains   67 Clusters
Maximum cluster size is 111 records
Minimum cluster size is  15 records

Independence parameters have converged in 5 iterations

Sample and Population Counts for Response Variable CANCER12
  Cancer   :  Sample Count      232    Population Count   1745695
  No Cancer:  Sample Count     3058    Population Count  38824628
```

REFLEVEL Example

Using the REFLEVEL Statement

```
Date: 05-29-97            Research Triangle Institute            Page  : 1
Time: 14:16:21              The MULTILOG Procedure               Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable CANCER12: Cancer Status (1/2)

Using the REFLEVEL Statement

----------------------------------------------------------------------
Independent Variables                              DESIGN
  and Effects                           BETA     S.E. EFFECT T:BETA=0  P-VALUE
----------------------------------------------------------------------
Intercept                            -3.7002   0.6967   1.06    -5.31   0.0000
Total Iron-Binding Capacity          -0.0024   0.0018   1.10    -1.29   0.2052
Age Cohort
  20-49 yrs.                          0.0000   0.0000    .        .       .
  50+ yrs.                            2.2525   0.3343   1.89     6.74   0.0000
Smoking Status
  Current                             0.0000   0.0000    .        .       .
  Former                             -0.3560   0.2716   1.16    -1.31   0.1985
  Never                               0.0860   0.2500   1.26     0.34   0.7330
  Unknown                             0.5858   0.2771   0.77     2.11   0.0417
----------------------------------------------------------------------
```

Now each smoking group is compared to the *current* smokers (SMOKE=1), and we see immediately that *current smokers* are not significantly different from *former smokers* (p=0.1985) nor from those who have *never smoked* (p=0.7330).

REFLEVEL Example

Using the REFLEVEL Statement

```
Date: 05-29-97             Research Triangle Institute             Page  : 2
Time: 14:16:21               The MULTILOG Procedure                Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable CANCER12: Cancer Status (1/2)

Using the REFLEVEL Statement

----------------------------------------------------
Contrast                  Degrees           P-value
                          of       Wald     Wald
                          Freedom  ChiSq    ChiSq
----------------------------------------------------
OVERALL MODEL                   6   708.28   0.0000
MODEL MINUS INTERCEPT           5    64.47   0.0000
B_TIBC                          1     1.67   0.1967
AGEGROUP                        1    45.39   0.0000
SMOKE                           3    10.60   0.0141
----------------------------------------------------

MULTILOG used
  CPU time       : 13.2 seconds
  Elapsed time   : 14 seconds
  Virtual memory : 2.88 MB
```

The tests of main effects are the same, no matter which groups are designated as the reference cells.

EFFECTS Statement

- Available in all modeling procedures

Simplifies the following hypothesis testing situations:

- Testing multiple main effects and/or interactions simultaneously (*e.g.*, testing chunk interaction effects);

- Testing general linear contrasts (*e.g.*, pairwise comparisons, trends) for a specific variable(s) in the model by only specifying contrast coefficients for the variable(s) of interest;

- Testing main effects in the presence of interactions. If the model contains factors A, B, and their interaction A*B, the user can obtain the:

  1) ***Simple effect*** of A, which is the effect of variable A tested within a given level of variable B, and
  2) ***Main effects*** of A, which are averaged over the levels of B.

Syntax:

```
EFFECTS term(s) / [ NAME = "label" ] [ DISPLAY ]
                  [ REFLEVEL | AVERAGE |
                    VARIABLE_NAME = value ] ;
```

where ***term(s)*** are name of effect(s) (single variables or/and interactions) on the MODEL statement, which may include contrast matrices.

EFFECTS Statement Options

NAME = ”*label*”

Assigns a label to the contrast. Default is “*Effect_nn*”, where *nn* is the *nn*-th EFFECT statement in the procedure

DISPLAY

Prints the contrast coefficients

REFLEVEL, AVERAGE , VARIABLE_NAME = *value*

Tells SUDAAN how to test the effects of covariates in the model when they are interacted with other effects in the model.

Example:

```
MODEL  Y  =  A   B  A*B;
```

To test the effect of A (which may be either continuous or categorical), the user has three options:

REFLEVEL (default)

Tests the effect of A ***when B (and all other variables A is interacted with) are set to their reference levels***.

AVERAGE

Tests the effect of A ***averaged over the interaction effect***, with proportional weighting over each level of B (Graubard and Korn, 1997). The contrast coefficient vector contains the weighted proportion of subjects in the *j*-th category of the *i*-th SUBGROUP variable.

EFFECTS Statement Options

VARIABLE_NAME = *value*

Similar to the REFLEVEL option, except here ***the user chooses the level of B within which to test the effect of A***. This option is used to carry out what are commonly known as “simple effects,” in which an effect A is to be tested within a specific level of B, other than the reference cell.

EFFECTS Example 1.

Using the NHANES I Study and its longitudinal follow-up (see the REFLEVEL statement examples for details), we evaluate the effects of body iron stores at initial exam (***B_TIBC***, continuous), age group at initial exam (***AGEGROUP***, 1=*20-49*, 2=*50+*), and smoking status (***SMOKE***, 1=*current*, 2=*former*, 3=*never*, 4=*unknown*) on follow-up cancer status (***CANCER12***, 1=*yes*, 2=*no*).

The **EFFECTS statement** can be used to:

1) Test the combined effect of *Agegroup* and *Smoke*:

```
EFFECTS AGEGROUP SMOKE /
        NAME = "Combined Age, Smoke";
```

2) Compare *Smoke* Level 1 to Level 2 (the default reference level for *Smoke* is Level 4):

```
EFFECTS SMOKE = (-1 1 0 0) / NAME="Smoke 1 vs 2";
```

EFFECTS Example 1.

```
1   PROC MULTILOG DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

2   NEST Q_STRATA PSU1;

3   WEIGHT B_WTIRON;

4   SUBGROUP CANCER12 AGEGROUP SMOKE;

5   LEVELS   2        2        4;

6   MODEL CANCER12 = B_TIBC AGEGROUP SMOKE / CUMLOGIT;

7   EFFECTS AGEGROUP SMOKE / NAME = "Combined Age, Smoke";

8   EFFECTS SMOKE=(-1 1 0 0) / NAME = "Smoke 1 vs 2";

9   SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

10  PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" DF WALDCHI WALDCHP /
          T_BETAFMT=F8.2 DEFTFMT=F6.2 DFFMT=F8.0 WALDCHIFMT=F8.2;

11  TITLE "EFFECTS Statement Example";


NOTE: Terms in the MODEL statement have been rearranged
      to follow subgroup order.

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.

Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is  6

File C:\ADVANCED\IRONSUD.SSD contains   67 Clusters
Maximum cluster size is 111 records
Minimum cluster size is  15 records

Independence parameters have converged in 5 iterations

Sample and Population Counts for Response Variable CANCER12
  Cancer   :  Sample Count      232    Population Count   1745695
  No Cancer:  Sample Count     3058    Population Count  38824628
```

EFFECTS Example 1.

```
Date: 05-29-97              Research Triangle Institute             Page  : 1
Time: 14:46:25                The MULTILOG Procedure                Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable CANCER12: Cancer Status (1/2)

EFFECTS Statement Example

----------------------------------------------------------------------
Independent Variables                            DESIGN
  and Effects                       BETA     S.E. EFFECT T:BETA=0  P-VALUE
----------------------------------------------------------------------
Intercept                        -0.8618   0.6605   0.94    -1.30   0.2004
Age Cohort
  20-49 yrs.                     -2.2525   0.3343   1.89    -6.74   0.0000
  50+ yrs.                        0.0000   0.0000    .        .       .
Smoking Status
  Current                        -0.5858   0.2771   0.77    -2.11   0.0417
  Former                         -0.9418   0.2922   0.84    -3.22   0.0027
  Never                          -0.4998   0.2743   0.85    -1.82   0.0770
  Unknown                         0.0000   0.0000    .        .       .
Total Iron-Binding Capacity      -0.0024   0.0018   1.10    -1.29   0.2052
----------------------------------------------------------------------
```

EFFECTS Example 1.

```
Date: 05-29-97              Research Triangle Institute             Page  : 2
Time: 14:46:25                The MULTILOG Procedure                Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable CANCER12: Cancer Status (1/2)

EFFECTS Statement Example

----------------------------------------------------
Contrast                  Degrees            P-value
                          of       Wald      Wald
                          Freedom  ChiSq     ChiSq
----------------------------------------------------
OVERALL MODEL                   6   708.28   0.0000
MODEL MINUS INTERCEPT           5    64.47   0.0000
AGEGROUP                        1    45.39   0.0000
SMOKE                           3    10.60   0.0141
B_TIBC                          1     1.67   0.1967
Combined Age, Smoke             4    53.16   0.0000
Smoke 1 vs 2                    1     1.72   0.1899
----------------------------------------------------

MULTILOG used
  CPU time       : 17.42 seconds
  Elapsed time   : 18 seconds
  Virtual memory : 2.88 MB
```

The combined effect of *Age* and *Smoking Status* is statistically significant (*p*=0.0000). However, *current smokers* (SMOKE=1) are not significantly different (*p*=0.1899) from *former smokers* (SMOKE=2).

EFFECTS Example 2.

In this example, we evaluate the effects of body iron stores at initial exam (***TRFSAT***, 1= *high* vs. 0=*normal* indicator), smoking status (***SMOKE***, 1=*current*, 2=*former*, 3=*never*, 4=*unknown*), age group at initial exam (***AGEGROUP,*** 1=*20-49* yrs, 2=*50+ yrs*), and various two-way interactions on a binary response, cancer status at follow-up (***CANCER1***, 1=*yes* vs. 0=*no*).

The **EFFECTS Statement** can be used to easily test simultaneous interaction effects (smoking by age group, smoking by indicator of body iron stores):

```
EFFECTS SMOKE*AGEGROUP SMOKE*TRFSAT / NAME="Chunk Interactions";
```

```
PROC LOGISTIC DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

NEST Q_STRATA PSU1;

WEIGHT B_WTIRON;

SUBGROUP SMOKE AGEGROUP;

LEVELS   4     2;

MODEL CANCER1 = TRFSAT SMOKE AGEGROUP SMOKE*AGEGROUP SMOKE*TRFSAT;

EFFECTS SMOKE*AGEGROUP SMOKE*TRFSAT / NAME = "Chunk Interactions";

SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
      P_BETA="P-VALUE" DF WALDCHI WALDCHP
      / SEBETAFMT=F8.5 DFFMT=F8.0 T_BETAFMT=F8.2 DEFTFMT=F6.2 WALDCHIFMT=F8.2;

TITLE "Using EFFECTS to Test Chunk Interactions";


NOTE: Terms in the MODEL statement have been rearranged
      to follow subgroup order.

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.
Number of zero responses     :  3058
Number of non-zero responses :   232

Parameters have converged in 5 iterations

Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is 12

R-Square for dependent variable CANCER1 (Cox & Snell, 1989): 0.046486
```

EFFECTS Example 2.

Using EFFECTS to Test Chunk Interactions

```
Date: 04-04-97            Research Triangle Institute            Page  : 1
Time: 15:55:41              The LOGISTIC Procedure               Table : 1

Response variable CANCER1: Cancer Status (0/1)

Using Effects to Test Chunk Interactions

--------------------------------------------------------------------
Independent Variables and
  Effects                                  DESIGN
                              BETA     S.E. EFFECT T:BETA=0  P-VALUE
--------------------------------------------------------------------
Intercept                  -1.6135  0.27254   0.72    -5.92   0.0000
Smoking Status
  Current                  -0.6159  0.37457   0.97    -1.64   0.1090
  Former                   -1.6133  0.33255   0.65    -4.85   0.0000
  Never                    -0.5606  0.35346   0.93    -1.59   0.1217
  Unknown                   0.0000  0.00000   .        .       .
Age Cohort
  20-49 yrs.               -3.8676  0.84072   0.31    -4.60   0.0001
  50+ yrs.                  0.0000  0.00000   .        .       .
High Transferrin
  Saturation (0/1)          0.1745  0.52386   0.72     0.33   0.7411
Smoking Status, Age Cohort
  Current, 20-49 yrs.       1.4407  1.03113   0.41     1.40   0.1711
  Current, 50+ yrs.         0.0000  0.00000   .        .       .
  Former, 20-49 yrs.        2.2305  1.05117   0.44     2.12   0.0410
  Former, 50+ yrs.          0.0000  0.00000   .        .       .
  Never, 20-49 yrs.         1.5366  1.03999   0.44     1.48   0.1485
  Never, 50+ yrs.           0.0000  0.00000   .        .       .
  Unknown, 20-49 yrs.       0.0000  0.00000   .        .       .
  Unknown, 50+ yrs.         0.0000  0.00000   .        .       .
Smoking Status, High
  Transferrin Saturation
  Current                  -0.1905  0.56612   0.58    -0.34   0.7385
  Former                    1.1955  0.69445   0.94     1.72   0.0940
  Never                    -0.1575  0.50445   0.52    -0.31   0.7568
  Unknown                   0.0000  0.00000   .        .       .
--------------------------------------------------------------------
```

EFFECTS Example 2.

Using EFFECTS to Test Chunk Interactions

```
Date: 04-04-97             Research Triangle Institute             Page  : 2
Time: 15:55:41               The LOGISTIC Procedure                Table : 1

Response variable CANCER1: Cancer Status (0/1)

Using EFFECTS to Test Chunk Interactions

----------------------------------------------------
Contrast                  Degrees           P-value
                          of       Wald     Wald
                          Freedom  ChiSq    ChiSq
----------------------------------------------------
OVERALL MODEL                   12   819.25   0.0000
MODEL MINUS INTERCEPT           11   101.61   0.0000
INTERCEPT                        .        .       .
SMOKE                            .        .       .
AGEGROUP                         .        .       .
TRFSAT                           .        .       .
SMOKE * AGEGROUP                 3     4.96   0.1749
TRFSAT * SMOKE                   3     6.02   0.1105
Chunk Interactions               6    21.21   0.0017
----------------------------------------------------
```

The combined interaction effect is statistically significant (p=0.0017). To test the same hypothesis using the CONTRAST statement, we would specify the following 12-row contrast matrix. The number of rows equals the number of regression coefficients to be tested in the contrast, with 1's in the columns corresponding to those regression coefficients. All other columns for intercept and main effects are 0's.

```
CONTRAST 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
         0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
         0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
         0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
         0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
         0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
         0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
         0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
         0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
         0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
         0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
         0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
         / NAME="CHUNK INTERACTIONS";
```

EFFECTS Example 2.

Comparison to the CONTRAST Statement

```
62  PROC LOGISTIC DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

63  NEST Q_STRATA PSU1;

64  WEIGHT B_WTIRON;

65  SUBGROUP SMOKE AGEGROUP;

66  LEVELS   4     2;

67  MODEL CANCER1=TRFSAT SMOKE AGEGROUP SMOKE*AGEGROUP SMOKE*TRFSAT;

68  CONTRAST 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
             0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
             0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
             0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
             0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
             0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
             0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
             0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
             0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
             0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
             0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
             0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
             / NAME="CHUNK INTERACTIONS";

69  SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

70  PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" DF WALDCHI WALDCHP / SEBETAFMT=F8.5 T_BETAFMT=F8.2
          DEFTFMT=F6.2 WALDCHIFMT=F8.2 DFFMT=F8.0;

71  TITLE "Using CONTRAST to Test Chunk Interactions";

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.
Number of zero responses     :  3058
Number of non-zero responses :   232

Parameters have converged in 5 iterations
Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is 12

R-Square for dependent variable CANCER1 (Cox & Snell, 1989): 0.046486
```

EFFECTS Example 2.

Comparison to the CONTRAST Statement

```
Date: 03-27-97            Research Triangle Institute            Page  : 1
Time: 14:25:00              The LOGISTIC Procedure               Table : 1

Response variable CANCER1: Cancer Status (0/1)
```

Using CONTRAST to Test Chunk Interactions

Independent Variables and Effects	BETA	S.E.	DESIGN EFFECT	T:BETA=0	P-VALUE
Intercept	-1.6135	0.27254	0.72	-5.92	0.0000
High Transferrin Saturation (0/1)	0.1745	0.52386	0.72	0.33	0.7411
Smoking Status					
Current	-0.6159	0.37457	0.97	-1.64	0.1090
Former	-1.6133	0.33255	0.65	-4.85	0.0000
Never	-0.5606	0.35346	0.93	-1.59	0.1217
Unknown	0.0000	0.00000	.	.	.
Age Cohort					
20-49 yrs.	-3.8676	0.84072	0.31	-4.60	0.0001
50+ yrs.	0.0000	0.00000	.	.	.
Smoking Status, Age Cohort					
Current, 20-49 yrs.	1.4407	1.03113	0.41	1.40	0.1711
Current, 50+ yrs.	0.0000	0.00000	.	.	.
Former, 20-49 yrs.	2.2305	1.05117	0.44	2.12	0.0410
Former, 50+ yrs.	0.0000	0.00000	.	.	.
Never, 20-49 yrs.	1.5366	1.03999	0.44	1.48	0.1485
Never, 50+ yrs.	0.0000	0.00000	.	.	.
Unknown, 20-49 yrs.	0.0000	0.00000	.	.	.
Unknown, 50+ yrs.	0.0000	0.00000	.	.	.
Smoking Status, High Transferrin Saturation (0/1)					
Current	-0.1905	0.56612	0.58	-0.34	0.7385
Former	1.1955	0.69445	0.94	1.72	0.0940
Never	-0.1575	0.50445	0.52	-0.31	0.7568
Unknown	0.0000	0.00000	.	.	.

EFFECTS Example 2.

Comparison to the CONTRAST Statement

```
Date: 03-27-97            Research Triangle Institute            Page  : 2
Time: 14:25:00              The LOGISTIC Procedure               Table : 1

Response variable CANCER1: Cancer Status (0/1)

Using CONTRAST to Test Chunk Interactions

----------------------------------------------------
Contrast                  Degrees            P-value
                          of        Wald     Wald
                          Freedom   ChiSq    ChiSq
----------------------------------------------------
OVERALL MODEL                   12   819.25   0.0000
MODEL MINUS INTERCEPT           11   101.61   0.0000
INTERCEPT                        .       .      .
TRFSAT                           .       .      .
SMOKE                            .       .      .
AGEGROUP                         .       .      .
SMOKE * AGEGROUP                 3     4.96   0.1749
TRFSAT * SMOKE                   3     6.02   0.1105
CHUNK INTERACTIONS               6    21.21   0.0017
----------------------------------------------------

LOGISTIC used
  CPU time       : 29.27 seconds
  Elapsed time   : 30 seconds
  Virtual memory : 2.23 MB
```

The results are the same as for the EFFECTS statement, with the simultaneous interactions being statistically significant.

EFFECTS Example 3.

In this example, we evaluate the effect of smoking status (***SMOKE***, 1=*current*, 2=*former*, 3=*never*, 4=*unknown*) on a binary response, cancer status at follow-up (***CANCER1***, 1=*yes* vs. 0=*no*) under the following conditions:

1) When Age Group=1 (20-49 yrs),
2) When Age Group=2 (50+ yrs),
3) When Age Group is at its reference level (level 2=50+ yrs),
4) Averaged over the interaction cells with Age Group.

The **EFFECTS statement** can be used to easily test these hypotheses:

```
EFFECTS SMOKE / AGEGROUP=1  NAME = "SMOKE in AGEGROUP=1";
EFFECTS SMOKE / AGEGROUP=2  NAME = "SMOKE in AGEGROUP=2";
EFFECTS SMOKE / REFLEVEL  NAME = "SMOKE in Age Reference Level";
EFFECTS SMOKE / AVERAGE  NAME = "SMOKE Averaged Over
                                 Interaction";
```

EFFECTS Example 3.

```
76  PROC LOGISTIC DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

77  NEST Q_STRATA PSU1;

78  WEIGHT B_WTIRON;

79  SUBGROUP AGEGROUP SMOKE;

80  LEVELS   2        4;

81  MODEL CANCER1 = TRFSAT AGEGROUP SMOKE AGEGROUP*SMOKE;

82  EFFECTS SMOKE / AGEGROUP=1 NAME="Smoke Effect in Age=20-49";

83  EFFECTS SMOKE / AGEGROUP=2 NAME="Smoke Effect in Age=50+";

84  EFFECTS SMOKE / REFLEVEL NAME="Smoke Effect at Age Reference Level";

85  EFFECTS SMOKE / AVERAGE NAME="Smoke averaged over interaction";

86  SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

87  PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" DF WALDCHI WALDCHP
          /SEBETAFMT=F8.5 DFFMT=F8.0 T_BETAFMT=F8.2 DEFTFMT=F6.2 WALDCHIFMT=F8.2;

88  TITLE "Using EFFECTS to Test Simple Effects;


NOTE: Terms in the MODEL statement have been rearranged
      to follow subgroup order.

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.
Number of zero responses     :  3058
Number of non-zero responses :   232

Parameters have converged in 5 iterations

Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is  9

R-Square for dependent variable CANCER1 (Cox & Snell, 1989): 0.043642
```

EFFECTS Example 3.

```
Date: 04-04-97            Research Triangle Institute             Page  : 1
Time: 15:55:41              The LOGISTIC Procedure                Table : 1

Response variable CANCER1: Cancer Status (0/1)

Using EFFECTS to Test Simple Effects

--------------------------------------------------------------------
Independent Variables and
  Effects                                  DESIGN
                              BETA     S.E. EFFECT T:BETA=0  P-VALUE
--------------------------------------------------------------------
Intercept                  -1.6762  0.25187   0.79    -6.65   0.0000
Age Cohort
  20-49 yrs.               -3.8681  0.84493   0.31    -4.58   0.0001
  50+ yrs.                  0.0000  0.00000    .        .       .
Smoking Status
  Current                  -0.6625  0.31953   0.91    -2.07   0.0455
  Former                   -1.1591  0.34790   1.05    -3.33   0.0020
  Never                    -0.6030  0.30426   0.91    -1.98   0.0554
  Unknown                   0.0000  0.00000    .        .       .
High Transferrin
  Saturation (0/1)          0.3997  0.20980   1.19     1.91   0.0650
Age Cohort, Smoking Status
  20-49 yrs., Current       1.4290  1.03443   0.41     1.38   0.1759
  20-49 yrs., Former        2.2399  1.04173   0.43     2.15   0.0385
  20-49 yrs., Never         1.5345  1.04652   0.45     1.47   0.1515
  20-49 yrs., Unknown       0.0000  0.00000    .        .       .
  50+ yrs., Current         0.0000  0.00000    .        .       .
  50+ yrs., Former          0.0000  0.00000    .        .       .
  50+ yrs., Never           0.0000  0.00000    .        .       .
  50+ yrs., Unknown         0.0000  0.00000    .        .       .
--------------------------------------------------------------------
```

EFFECTS Example 3.

```
Date: 04-04-97           Research Triangle Institute            Page  : 2
Time: 15:55:41             The LOGISTIC Procedure               Table : 1

Response variable CANCER1: Cancer Status (0/1)

Using EFFECTS to Test Simple Effects

--------------------------------------------------------
Contrast                       Degrees           P-value
                               of        Wald    Wald
                               Freedom   ChiSq   ChiSq
--------------------------------------------------------
OVERALL MODEL                        9   859.59   0.0000
MODEL MINUS INTERCEPT                8    89.15   0.0000
INTERCEPT                            .      .       .
AGEGROUP                             .      .       .
SMOKE                                .      .       .
TRFSAT                               1     3.63   0.0567
AGEGROUP * SMOKE                     3     5.25   0.1547
Smoke Effect in Age=20-49            3     1.66   0.6466
Smoke Effect in Age=50+              3    11.15   0.0110
Smoke Effect at Age Reference Level  3    11.15   0.0110
Smoke averaged over interaction      3     0.36   0.9491
--------------------------------------------------------

LOGISTIC used
  CPU time       : 25.87 seconds
  Elapsed time   : 26 seconds
  Virtual memory : 2.02 MB
```

Note that the test for *"Smoke Effect in Age=50+"* is equivalent to *"Smoke in Age Reference Level."* Here we see that:

1) There is a marginally significant interaction between age and smoking on follow-up cancer status (p=0.1547). SUDAAN computes this test automatically, without the need for the EFFECTS statement.

2) There is no significant effect of smoking on cancer status when age group=20-49 yrs. (p=0.6466), although the regression coefficients on the previous page (provided automatically by SUDAAN) and the EFFECTS statement here indicates a significant smoking effect when age is at its reference level (50+ yrs., p=0.0110).

3) There is no significant effect of smoking when smoking is averaged over its interaction with age (p=0.9302).

Now the same results via the CONTRAST statement:

EFFECTS Example 3.

Comparison to the CONTRAST Statement

```
72  PROC LOGISTIC DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

73  NEST Q_STRATA PSU1;

74  WEIGHT B_WTIRON;

75  SUBGROUP AGEGROUP SMOKE;

76  LEVELS   2        4;

77  MODEL CANCER1 = TRFSAT AGEGROUP SMOKE AGEGROUP*SMOKE;

78  CONTRAST 0 0 0 0 -1 0 0 1 -1 0 0 1 0 0 0 0
             0 0 0 0 -1 0 1 0 -1 0 1 0 0 0 0 0
             0 0 0 0 -1 1 0 0 -1 1 0 0 0 0 0 0
             / NAME="SMOKE IN AGE=1";

79  CONTRAST 0 0 0 0 -1 0 0 1 0 0 0 0 -1 0 0 1
             0 0 0 0 -1 0 1 0 0 0 0 0 -1 0 1 0
             0 0 0 0 -1 1 0 0 0 0 0 0 -1 1 0 0
             / NAME="SMOKE IN AGE=2";

80  SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

81  PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
          P_BETA="P-VALUE" DF WALDCHI WALDCHP / SEBETAFMT=F8.5 DFFMT=F8.0
          T_BETAFMT=F8.2 DEFTFMT=F6.2 WALDCHIFMT=F8.2;

82  TITLE "Testing Simple Effects via the CONTRAST Statement";

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.
Number of zero responses     :  3058
Number of non-zero responses :   232

Parameters have converged in 5 iterations

Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is  9

R-Square for dependent variable CANCER1 (Cox & Snell, 1989): 0.043642
```

EFFECTS Example 3.

Comparison to the CONTRAST Statement

```
Date: 03-27-97             Research Triangle Institute            Page  : 1
Time: 14:25:00               The LOGISTIC Procedure               Table : 1

Response variable CANCER1: Cancer Status (0/1)
```

Testing Simple Effects Via the CONTRAST Statement

Independent Variables and Effects	BETA	S.E.	DESIGN EFFECT	T:BETA=0	P-VALUE
Intercept	-1.6762	0.25187	0.79	-6.65	0.0000
High Transferrin Saturation (0/1)	0.3997	0.20980	1.19	1.91	0.0650
Age Cohort					
20-49 yrs.	-3.8681	0.84493	0.31	-4.58	0.0001
50+ yrs.	0.0000	0.00000	.	.	.
Smoking Status					
Current	-0.6625	0.31953	0.91	-2.07	0.0455
Former	-1.1591	0.34790	1.05	-3.33	0.0020
Never	-0.6030	0.30426	0.91	-1.98	0.0554
Unknown	0.0000	0.00000	.	.	.
Age Cohort, Smoking Status					
20-49 yrs., Current	1.4290	1.03443	0.41	1.38	0.1759
20-49 yrs., Former	2.2399	1.04173	0.43	2.15	0.0385
20-49 yrs., Never	1.5345	1.04652	0.45	1.47	0.1515
20-49 yrs., Unknown	0.0000	0.00000	.	.	.
50+ yrs., Current	0.0000	0.00000	.	.	.
50+ yrs., Former	0.0000	0.00000	.	.	.
50+ yrs., Never	0.0000	0.00000	.	.	.
50+ yrs., Unknown	0.0000	0.00000	.	.	.

EFFECTS Example 3.

Comparison to the CONTRAST Statement

```
Date: 03-27-97             Research Triangle Institute             Page  : 2
Time: 14:25:00               The LOGISTIC Procedure                Table : 1

Response variable CANCER1: Cancer Status (0/1)

Testing Simple Effects Via the CONTRAST Statement

----------------------------------------------------
Contrast                   Degrees           P-value
                           of       Wald     Wald
                           Freedom  ChiSq    ChiSq
----------------------------------------------------
OVERALL MODEL                    9   859.59   0.0000
MODEL MINUS INTERCEPT            8    89.15   0.0000
INTERCEPT                        .      .       .
TRFSAT                           1     3.63   0.0567
AGEGROUP                         .      .       .
SMOKE                            .      .       .
AGEGROUP * SMOKE                 3     5.25   0.1547
SMOKE IN AGE=1                   3     1.66   0.6466
SMOKE IN AGE=2                   3    11.15   0.0110
----------------------------------------------------

LOGISTIC used
  CPU time       : 23.95 seconds
  Elapsed time   : 24 seconds
  Virtual memory : 2.07 MB
```

LSMEANS Statement

- Available in the ***linear regression procedure*** (REGRESS).

- Produces “least squares” or “adjusted means” for any number of categorical covariates in the model.

- List one or more categorical effects from the right-hand-side of the MODEL statement. ***Continuous variables are not allowed*** on the LSMEANS statement.

- The keyword ***INTERCEPT*** specifies an overall least-squares mean, when the model contains an intercept.

Syntax:

```
LSMEANS [INTERCEPT] effect(s) / [ALL] [DISPLAY] ;
```

ALL

Requests least-squares means for ***all effects*** on the right-hand side of the MODEL statement.

DISPLAY

Requests least squares means ***contrast coefficients***.

LSMEANS Statement

Construction of the LSMEANS Contrast

- SUDAAN calculates ***contrast coefficients*** that are the weighted means of each covariate to be adjusted for in the model, using all observations for which there are no missing independent or dependent variable values.

- Contrast coefficients corresponding to the levels of the ***categorical covariates*** (appearing on the SUBGROUP statement) are the weighted numbers of individuals in each category of the covariate. Sample member weights are provided by the variable specified on the WEIGHT statement. If weights are all equal to one (*e.g.*, via the keyword _ONE_), unweighted means are used.

- The set of contrast coefficients are vector-multiplied by the estimated regression coefficients.

LSMEANS Example

The following example illustrates the construction of the LSMEANS contrast.

Data:
NHANES I Survey and its Longitudinal Follow-up Study.

Question:
Is smoking status at initial exam (***SMOKE***, where 1=*current* vs. 2=*former*, 3=*never*, 4=*unknown*) associated with a measure of body iron stores at the initial exam (***B_TIBC***, or total iron-binding capacity), while adjusting for age at initial exam?

LSMEANS
We request the least squares means of the response *B_TIBC*, total iron-binding capacity, within levels of *SMOKE*, adjusted for age at initial exam (first as categorical, then as a continuous covariate). The data are weighted by the variable *B_WTIRON*.

SUDAAN Programming Statements Demonstrating the Construction of the LSMEANS Contrast for Categorical Covariates

```
PROC REGRESS DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

NEST Q_STRATA PSU1;

WEIGHT B_WTIRON;

SUBGROUP AGEGROUP SMOKE;

LEVELS   2        4;

MODEL B_TIBC = SMOKE AGEGROUP;

LSMEANS SMOKE / DISPLAY;

SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
      P_BETA="P-VALUE" DF WALDCHI WALDCHP /
      LSMEANS=ALL T_BETAFMT=F8.2 DEFTFMT=F6.2 DFFMT=F8.0 WALDCHIFMT=F8.2;

TITLE "LSMEANS With Categorical Covariate";

NOTE: Terms in the MODEL statement have been rearranged
      to follow subgroup order.

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.

Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is  5

File C:\ADVANCED\IRONSUD.SSD contains   67 clusters
Maximum cluster size is 111 records
Minimum cluster size is  15 records
Weighted mean response is 354.580621
```

LSMEANS Example

Estimated Regression Coefficients for the Model

```
Date: 05-29-97            Research Triangle Institute            Page  : 4
Time: 15:28:17               The REGRESS Procedure               Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable B_TIBC: TOTAL IRON BINDING CAPACITY

LSMEANS With Categorical Covariate

--------------------------------------------------------------------
Independent Variables and
  Effects                                   DESIGN
                              BETA     S.E. EFFECT T:BETA=0  P-VALUE
--------------------------------------------------------------------
Intercept                 352.8876   3.8547   1.09    91.55   0.0000
Age Cohort
  20-49 yrs.                7.2210   1.8968   1.12     3.81   0.0005
  50+ yrs.                  0.0000   0.0000    .        .       .
Smoking Status
  Current                  -7.5062   3.7690   0.95    -1.99   0.0543
  Former                   -1.6754   4.2636   1.25    -0.39   0.6967
  Never                    -0.9261   3.8284   1.06    -0.24   0.8103
  Unknown                   0.0000   0.0000    .        .       .
--------------------------------------------------------------------
```

LSMEANS Example

Least Squares Means Contrast Coefficients:

Smoking Status and Age Group

Since we want to estimate the least squares means of the response within each level of smoking status (a 4-level variable), SUDAAN will produce four rows of contrast coefficients. The first row of the matrix will produce the adjusted means for SMOKE=*current*, the second row is for SMOKE=*former*, and so on. The contrast coefficients for ***smoking status*** are 1's and 0's, indicating the level of interest. Since we are adjusting for ***age group*** as a categorical covariate, the age group coefficients are the weighted (weight = *b_wtiron*) proportion of people in each of the two categories.

Age Group Contrast Coefficients

```
Date: 05-29-97           Research Triangle Institute              Page  : 1
Time: 15:28:17              The REGRESS Procedure                 Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable B_TIBC: TOTAL IRON BINDING CAPACITY

LS Means Contrast

----------------------------------------------------------------------
                                              Age Cohort      Age Cohort
                             Intercept       20-49 yrs.        50+ yrs.
----------------------------------------------------------------------
Smoking Status
  Current                        1.000           0.603           0.397
  Former                         1.000           0.603           0.397
  Never                          1.000           0.603           0.397
  Unknown                        1.000           0.603           0.397
----------------------------------------------------------------------
```

LSMEANS Example

Least Squares Means Contrast Coefficients:

Smoking Status Coefficients

The contrast coefficients for ***smoking status*** are 1's and 0's, indicating the level of interest in each row.

```
Date: 05-29-97            Research Triangle Institute              Page  : 2
Time: 15:28:17               The REGRESS Procedure                 Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable B_TIBC: TOTAL IRON BINDING CAPACITY

LS Means Contrast
```

	Smoking Status **Current**	Smoking Status **Former**	Smoking Status **Never**	Smoking Status **Unknown**
Smoking Status				
Current	**1.000**	0.000	0.000	0.000
Former	0.000	**1.000**	0.000	0.000
Never	0.000	0.000	**1.000**	0.000
Unknown	0.000	0.000	0.000	**1.000**

LSMEANS Example

Least Squares Means Results

Age Group as Categorical Covariate

This table shows the ***estimated least-squares means***, with standard errors that are adjusted for clustering and stratification (via the NEST statement and DESIGN=WR option on the PROC statement).

```
Date: 05-29-97            Research Triangle Institute            Page  : 6
Time: 15:28:17               The REGRESS Procedure               Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable B_TIBC: TOTAL IRON BINDING CAPACITY

LSMEANS With Categorical Covariate

--------------------------------------------------------------
Least-Square Means                                   P-value
                                    SE LS    T-Test    T-Test
                           LS Mean  Mean     LSM=0     LSM=0
--------------------------------------------------------------
Smoking Status
  Current                 349.7372  2.1938  159.4181   0.0000
  Former                  355.5680  2.2920  155.1367   0.0000
  Never                   356.3173  2.0476  174.0141   0.0000
  Unknown                 357.2434  3.5898   99.5154   0.0000
--------------------------------------------------------------
```

LSMEANS Example

Least Squares Means Contrast Coefficients:

Age at Exam as Continuous Covariate

Now we show how the contrast is formed when age is modelled as a ***continuous*** covariate.

```
PROC REGRESS DATA="C:\\ADVANCED\\IRONSUD" FILETYPE=SAS DESIGN=WR DEFT2;

NEST Q_STRATA PSU1;

WEIGHT B_WTIRON;

SUBGROUP SMOKE;

LEVELS   4;

MODEL B_TIBC = SMOKE AGEXAM;

LSMEANS SMOKE / DISPLAY;

SETENV COLSPCE=1 LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60;

PRINT BETA="BETA" SEBETA="S.E." DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
      P_BETA="P-VALUE" DF WALDCHI WALDCHP /
      LSMEANS=ALL T_BETAFMT=F8.2 DEFTFMT=F6.2 DFFMT=F8.0 WALDCHIFMT=F8.2;

TITLE "LSMEANS With Continuous Covariate";

Opened SAS data file C:\ADVANCED\IRONSUD.SSD for reading.

Number of observations read       :   3290    Weighted count: 40570323
Observations used in the analysis :   3290    Weighted count: 40570323
Observations with missing values  :      0    Weighted count:        0
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is  5

File C:\ADVANCED\IRONSUD.SSD contains   67 clusters
Maximum cluster size is 111 records
Minimum cluster size is  15 records
Weighted mean response is 354.580621
```

LSMEANS Example

Estimated Regression Coefficients for the Model

Age at Exam as Continuous Covariate

```
Date: 05-29-97            Research Triangle Institute            Page  : 3
Time: 15:28:17               The REGRESS Procedure               Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable B_TIBC: TOTAL IRON BINDING CAPACITY

LSMEANS With Continuous Covariate

--------------------------------------------------------------------
Independent Variables and
  Effects                                    DESIGN
                              BETA     S.E.  EFFECT T:BETA=0  P-VALUE
--------------------------------------------------------------------
Intercept                 370.4372   4.9483   1.19    74.86   0.0000
Smoking Status
  Current                  -8.0845   3.7812   0.95    -2.14   0.0396
  Former                   -2.0617   4.2763   1.26    -0.48   0.6327
  Never                    -1.5183   3.8930   1.09    -0.39   0.6989
  Unknown                   0.0000   0.0000    .        .       .
Age at Exam                -0.2778   0.0730   1.27    -3.81   0.0005
--------------------------------------------------------------------
```

LSMEANS Example

Least Squares Means Contrast Coefficients:

Age at Exam as Continuous Covariate

When age at initial exam is modelled as a continuous covariate, its single contrast coefficient is the weighted mean of ***AGEXAM*** (45.706 years). The contrast coefficients for Smoking status are the same as previously.

```
Date: 05-29-97            Research Triangle Institute              Page  : 2
Time: 15:28:17               The REGRESS Procedure                 Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable B_TIBC: TOTAL IRON BINDING CAPACITY

LS Means Contrast

--------------------------------
                     Age at Exam
--------------------------------
Smoking Status
  Current                 45.706
  Former                  45.706
  Never                   45.706
  Unknown                 45.706
--------------------------------
```

LSMEANS Example

Least Squares Means Results with Age as Continuous Covariate

This table shows the ***estimated least-squares means***, with standard errors that are adjusted for clustering and stratification (via the NEST statement and DESIGN=WR option on the PROC statement), when Age is modelled as a continuous covariate.

```
Date: 05-29-97            Research Triangle Institute            Page  : 5
Time: 15:28:17               The REGRESS Procedure               Table : 1

Variance Estimation Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable B_TIBC: TOTAL IRON BINDING CAPACITY

LSMEANS With Continuous Covariate

---------------------------------------------------------------
Least-Square Means                                     P-value
                                     SE LS    T-Test    T-Test
                          LS Mean    Mean     LSM=0     LSM=0
---------------------------------------------------------------
Smoking Status
  Current                349.6539   2.2333   156.5668   0.0000
  Former                 355.6767   2.2900   155.3203   0.0000
  Never                  356.2201   2.0547   173.3643   0.0000
  Unknown                357.7384   3.6201    98.8206   0.0000
---------------------------------------------------------------
```

Design Effects

Numerator
Variance calculated according to the user-specified sample design option, the working correlation structure specified (independent or exchangeable), and the standard error method (robust vs. model-based).

Denominator = SRS Variance
Calculated according to the type of design effect requested on the PROC statement (see below: DEFT1, DEFT2, DEFT3, DEFT4). DEFT4 is the default, and leads to SRSCOV calculated as the model-based variance under the user-specified correlation structure (independent vs. exchangeable).

Design Effect	Measures Variance Inflation Due to:	Default?
DEFT1	Stratification (or blocking), Clustering, Unequal Weighting, and ***Oversampling*** Assumes that total sample size is fixed	No; This is the original one; Request on PROC Statement
DEFT2	Stratification (or blocking), Clustering, and Unequal Weighting Assumes that subgroup sample sizes are fized	No; Request on PROC statement
DEFT3	Stratification (or blocking), Clustering Assumes that subgroup sample sizes are fixed	No; Request on PROC Statement
DEFT4	Stratification (or blocking), Clustering, and Unequal Weighting: ***Model-based*** SRS variance (this is the standard software variance when no weights involved) ***Good for experimental designs***	**Yes**

Evaluation of a Drug Abuse Prevention Program: Project DARE

- Ennett, Rosenbaum, Flewelling, Bieler, *et al* (1994)
  Norton, Bieler, Ennett, and Zarkin (1996)

- Longitudinal evaluation of the DARE program (Drug Abuse Resistance Education) in northern and central Illinois

- Semester-long drug-use prevention program for upper elementary school students (5th and 6th graders)

- Convenience sample of 36 schools (clusters) representing urban, suburban, and rural areas, randomly assigned to DARE and control conditions

- Data represent responses from students immediately before and after program implementation (Waves 1 and 2)

- 1,525 students present at both waves of data collection

- Outcome: Initiation of cigarette smoking by Wave 2 (includes only those students reporting no lifetime use in wave 1)

$$y = \begin{cases} 0, & \text{if student did not initiate smoking by follow-up} \\ 1, & \text{if student initiated smoking by follow-up} \end{cases}$$

Question: **Does the DARE program reduce the incidence of adolescent cigarette smoking (at least during the intervention)?**

Fitting GEE Logistic Regression Models in MULTILOG

Evaluation of a Drug Abuse Prevention Program (Project DARE)

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). The DARE curriculum is a semester-long drug use prevention program for late elementary school students. Respondents for the study were originally obtained in 1990 from the fifth and sixth grades of 36 schools representative of rural, urban, and suburban areas in the state of Illinois. Within each metropolitan status stratum, 6 pairs of schools (matched on various demographic characteristics) were randomly assigned to DARE and control conditions.

Researchers collected data immediately before and after program implementation (Waves 1 and 2) and have collected three additional waves at annual intervals since then. Analyses reported here draw only on data from Waves 1 and 2. The sample includes students for whom complete information is available on the variables of interest in both waves (N = 1525, 85% of Wave 1 sample). Students answered a self-administered questionnaire that took approximately 35 minutes to complete. The questions concerned substance use, attitudes toward drugs, self-esteem, and peer-resistance skills.

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 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 SUDAANs 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 here 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 Table 1, we compare the GEE/SUDAAN results to SAS PROC LOGISTIC, which currently fits ordinary logistic regression but naively makes no correction for intracluster correlation and instead considers the observations statistically independent.

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 no matter 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 modelling the correlation structure (*e.g.*, through exchangeability) is its potential to improve efficiency and hence increase the power of statistical analyses.

The 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 Table 1, naively ignoring the intracluster correlation as in SAS PROC LOGISTIC 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.

Structure of the DARE Data

Exposure Group 1 = Control 2 = DARE	**School ID (Cluster)**	**Student ID** (unit of observation)	**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)

Evaluation of the DARE Effect on Cigarette Initiation Via Logistic Regression Modelling

		Working Correlations			
		Independent (Ordinary Logistic Regression)		**Exchangeable**	
Variable	**Statistic**	**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	--	1.77
	Z-statistic	-2.87	-2.17	-2.39	-2.41
	P-value	**0.0069**	**0.0369**	**0.0221**	**0.0216**

Working Correlations: ***Software:***

Independent
(Ordinary Logistic Regression)

No variance correction: SAS Logistic

Variance Correction
(robust variance): SUDAAN Multilog

Exchangeable

No variance correction
[model-based (naive) variance]: SUDAAN Multilog

Variance Correction
(robust variance): SUDAAN Multilog

MULTILOG Programming Statements and Options

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

In the accompanying output, we fit the following types of GEE logistic regression models:

1) ***SEMETHOD=ZEGER*** and ***R=INDEPENDENT***
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.

2) ***SEMETHOD=MODEL*** and ***R=INDEPENDENT***
This amounts to ordinary logistic regression without a variance correction, which yields the same results as SAS PROC LOGISTIC. 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.

3) ***SEMETHOD=ZEGER*** and ***R=EXCHANGEABLE***
Implements the GEE model-fitting technique under ***exchangeable "working" correlations*** and Zeger and Liang's (1986) ***robust*** variance estimator.

4) ***SEMETHOD=MODEL*** and ***R=EXCHANGEABLE***
We compare the results from the robust variance estimator (*SEMETHOD=ZEGER*) to the ***model-based***, or ***naive***, variance assumption (*SEMETHOD=MODEL*). When *R=exchangeable* is specified in conjunction with *SEMETHOD=MODEL,* variances are then computed as if the ***exchangeable "working" correlation*** assumption were correct.

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 **SUBGROUP** statement contains the dependent variable and all covariates that are to be modelled as categorical covariates (with level values of 1,2,...,*K*), where the maximum number of levels (*K*) appears on the **LEVELS** statement.

The **MODEL** statement specifies the categorical dependent variable INTCIG12 on the left of the "=" sign (with levels 1 and 2), and regressors on the right. For binary responses, the **CUMLOGIT** (cumulative logit) and **GENLOGIT** (generalized logit) links specify the same logistic regression model.

The **TEST** statement specifies that we want the Wald chi-square statistic to be the default for testing main effects, interactions, and user-defined contrasts.

Results

Descriptive Statistics for Initiation of Cigarette Smoking in the DARE Study

```
1   PROC DESCRIPT DATA="c:\\tera\\examples\\DARE" FILETYPE=SAS NOMARG;

2   NEST _ONE_ SCHOOL;

3   WEIGHT _ONE_;

4   SUBGROUP DARE;

5   LEVELS 2;

6   TABLES DARE;

7   VAR INTCIG12;

8   CATLEVEL 1;

9   SETENV LABWIDTH=30 COLWIDTH=6 DECWIDTH=2;

10  PRINT NSUM PERCENT SEPERCENT="STDERR" DEFFPCT="Design Effect" /
          NSUMFMT=F6.0 PERCENTFMT=F7.2 STYLE=NCHS;

11  TITLE "DESCRIPTIVE STATISTICS FOR THE DARE DATA";

Opened SAS data file c:\tera\examples\DARE.SSD for reading.
Number of observations read    :   1525    Weighted count :     1525
Denominator degrees of freedom :     35

Date: 04-28-97             Research Triangle Institute             Page  : 1
Time: 13:30:35                The DESCRIPT Procedure               Table : 1

Variance Estimation Method: Taylor Series (WR)

DESCRIPTIVE STATISTICS FOR THE DARE DATA
-------------------------------------------------------------------
Variable                         Sample                        Design
   DARE Program                  Size     Percent   STDERR   Effect
-------------------------------------------------------------------
Initiation of Cigarette Use: Yes
   Exposed to DARE                  649      9.55     1.77     2.35
   Not Exposed                      539     15.40     2.25     2.10
-------------------------------------------------------------------
```

These results indicate that 15.4% of students not receiving DARE initiated cigarette smoking during the time of the intervention, 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. Is the observed difference statistically significant, after adjustment for other covariates? The MULTILOG procedure will be used to find out.

Results:

GEE With Independent "Working" Correlations
Robust Variance Estimator

```
12  PROC MULTILOG DATA="c:\\tera\\examples\\DARE" FILETYPE=SAS
                   SEMETHOD=ZEGER R=INDEPENDENT;

13  NEST _ONE_ SCHOOL;

14  WEIGHT _ONE_;

15  SUBGROUP DARE FIFTH SEX RACE OTHFAM AREA INTCIG12;

16  LEVELS   2    2     2   4    2      3    2;

17  MODEL INTCIG12 = DARE FIFTH SEX RACE OTHFAM AREA / CUMLOGIT;

18  TEST WALDCHI;

19  SETENV LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60 COLSPCE=2;

20  PRINT  BETA="BETA" SEBETA="STDERR" DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
           P_BETA="P-VALUE" / RISK=ALL TESTS=DEFAULT
           BETAFMT=F8.6 SEBETAFMT=F8.6 T_BETAFMT=F8.2 WALDCHIFMT=F6.2
           WALDCHPFMT=F7.4 DEFTFMT=F6.2 DFFMT=F7.0
           ORFMT=F5.2 LOWORFMT=F6.2 UPORFMT=F6.2;

21  TITLE "MULTILOG Logistic Regression Model for the DARE Evaluation Study"
          "Ennett, et al, 1994";

Opened SAS data file c:\tera\examples\DARE.SSD for reading.

Number of observations read       :   1525    Weighted count:     1525
Observations used in the analysis :   1188    Weighted count:     1188
Observations with missing values  :    337    Weighted count:      337
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is 10
File c:\tera\examples\DARE.SSD contains   36 Clusters
Maximum cluster size is 153 records
Minimum cluster size is  11 records

Independence parameters have converged in 4 iterations
Sample and Population Counts for Response Variable INTCIG12
  Yes:  Sample Count      145    Population Count       145
  No :  Sample Count     1043    Population Count      1043
```

Here we see that there are 1,525 students (1 record/student) on the file, and that 1,188 were used in the analysis (337 students deleted due to missing values on one or more MODEL statement variables). There are 36 clusters (schools), with cluster sizes ranging from 11 to 153. Overall, 145 students reported having initiated cigarette use during the intervention, while 1043 did not.

Results:

GEE With Independent "Working" Correlations
Robust Variance Estimator

```
Date: 04-28-97             Research Triangle Institute             Page  : 1
Time: 13:30:35               The MULTILOG Procedure                Table : 1

Variance Estimation 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

-------------------------------------------------------------------------
Independent Variables                                 DESIGN
  and Effects                    BETA    STDERR  EFFECT  T:BETA=0   P-VALUE
-------------------------------------------------------------------------
INTCIG12 (cum-logit)
Intercept 1: Yes           -1.84755  0.465860    1.99     -3.97    0.0003
DARE Program
  Yes                      -0.52248  0.240770    1.75     -2.17    0.0369
  No                       0.000000  0.000000     .         .        .
Grade in School
  5th Grade                -0.50020  0.249380    1.25     -2.01    0.0527
  6th Grade                0.000000  0.000000     .         .        .
SEX
  Male                     0.084014  0.159940    0.78      0.53    0.6027
  Female                   0.000000  0.000000     .         .        .
RACE
  Black                    0.497135  0.378610    1.78      1.31    0.1977
  Hispanic                 0.095132  0.467026    1.46      0.20    0.8398
  Other                    0.493601  0.421419    2.23      1.17    0.2494
  White                    0.000000  0.000000     .         .        .
Family Situation
  Non-Traditional          0.420841  0.170648    0.78      2.47    0.0187
  Traditional              0.000000  0.000000     .         .        .
AREA
  Rural                    -0.07878  0.396184    1.53     -0.20    0.8435
  Suburban                 -0.25078  0.360992    1.77     -0.69    0.4918
  Urban                    0.000000  0.000000     .         .        .
-------------------------------------------------------------------------
```

This first table contains the estimated regression coefficient vector, the estimated robust standard errors, design effects, t-statistics, and p-values for testing H_0: $\beta=0$. The CUMLOGIT option estimates only one model intercept in the case of a binary outcome, and is equivalent to the GENLOGIT option. The treatment effect (DARE) is observed to significantly reduce the incidence of cigarette initiation (p=0.0369) using the GEE-independent approach, after adjusting for other covariates in the model. 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.

Results:

GEE With Independent "Working" Correlations
Robust Variance Estimator

```
Date: 04-28-97             Research Triangle Institute            Page  : 2
Time: 13:30:35               The MULTILOG Procedure               Table : 1

Variance Estimation 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          P-value
                           of       Wald    Wald
                           Freedom  ChiSq   ChiSq
---------------------------------------------------
OVERALL MODEL                   10  313.24   0.0000
MODEL MINUS INTERCEPT            9   28.27   0.0009
DARE                             1    4.71   0.0300
FIFTH                            1    4.02   0.0449
SEX                              1    0.28   0.5994
RACE                             3    1.90   0.5930
OTHFAM                           1    6.08   0.0137
AREA                             2    0.60   0.7420
---------------------------------------------------
```

This table contains the statistical significance of all main effects, interactions, and user-defined contrasts. The Wald chi-square test (from the TEST statement) is used to evaluate these effects.

Results:

GEE With Independent "Working" Correlations
Robust Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 3
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation 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

------------------------------------------------
Independent Variables                 Lower   Upper
  and Effects              Odds        95%     95%
                           Ratio      Limit   Limit
------------------------------------------------
INTCIG12 (cum-logit)
Intercept 1: Yes            0.16       0.06    0.41
DARE Program
  Yes                       0.59       0.36    0.97
  No                        1.00       1.00    1.00
Grade in School
  5th Grade                 0.61       0.37    1.01
  6th Grade                 1.00       1.00    1.00
SEX
  Male                      1.09       0.79    1.50
  Female                    1.00       1.00    1.00
RACE
  Black                     1.64       0.76    3.55
  Hispanic                  1.10       0.43    2.84
  Other                     1.64       0.70    3.85
  White                     1.00       1.00    1.00
Family Situation
  Non-Traditional           1.52       1.08    2.15
  Traditional               1.00       1.00    1.00
AREA
  Rural                     0.92       0.41    2.07
  Suburban                  0.78       0.37    1.62
  Urban                     1.00       1.00    1.00
------------------------------------------------

MULTILOG used
  CPU time       : 12.3 seconds
  Elapsed time   : 13 seconds
  Virtual memory : 1.52 MB
```

This final table contains 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 0.59, 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.

Results

GEE with Independent "Working" Correlations
Model-Based (Naive) Variance Estimator

Below are the results obtained under working independence using the ***model-based*** or ***naive variance-covariance matrix*** of the estimated regression coefficients. The model-based variance is the M_0^{-1} matrix, or the outside portion of the robust variance estimate: $M_0^{-1} = [\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}]^{-1}$, 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 are showing it to demonstrate the effects of clustering. We use the ***SEMETHOD=MODEL*** option on the PROC statement to obtain the model-based results.

```
22  PROC MULTILOG DATA="c:\\tera\\examples\\DARE" FILETYPE=SAS
                  SEMETHOD=MODEL R=INDEPENDENT;

23  NEST _ONE_ SCHOOL;

24  WEIGHT _ONE_;

25  SUBGROUP DARE FIFTH SEX RACE OTHFAM AREA INTCIG12;

26  LEVELS   2    2     2   4    2      3    2;

27  MODEL INTCIG12 = DARE FIFTH SEX RACE OTHFAM AREA / CUMLOGIT;

28  TEST WALDCHI;

29  SETENV LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 MAXIND=4 LINESIZE=78
           PAGESIZE=60 COLSPCE=2;

30  PRINT  BETA="BETA" SEBETA="STDERR" DEFT="DESIGN EFFECT" T_BETA="T:BETA=0"
           P_BETA="P-VALUE" / RISK=ALL TESTS=DEFAULT
           BETAFMT=F8.6 SEBETAFMT=F8.6 T_BETAFMT=F8.2 WALDCHIFMT=F6.2
           WALDCHPFMT=F7.4 DEFTFMT=F6.2 DFFMT=F7.0
           ORFMT=F5.2 LOWORFMT=F6.2 UPORFMT=F6.2;

31  TITLE "MULTILOG Logistic Regression Model for the DARE Evaluation Study"
          "Model-Based Variance Estimation";

Opened SAS data file c:\tera\examples\DARE.SSD for reading.

Number of observations read       :   1525    Weighted count:     1525
Observations used in the analysis :   1188    Weighted count:     1188
Observations with missing values  :    337    Weighted count:      337
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is 10
```

```
File c:\tera\examples\DARE.SSD contains   36 Clusters
Maximum cluster size is 153 records
Minimum cluster size is  11 records

Independence parameters have converged in 4 iterations

Sample and Population Counts for Response Variable INTCIG12
  Yes:  Sample Count      145    Population Count       145
  No :  Sample Count     1043    Population Count      1043
```

Results

GEE with Independent "Working" Correlations
Model-Based (Naive) Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 1
Time: 13:30:35              The MULTILOG Procedure               Table : 1
```

Variance Estimation Method: Model-Based (Naive)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Model-Based Variance Estimation

Independent Variables and Effects	BETA	STDERR	DESIGN EFFECT	T:BETA=0	P-VALUE
INTCIG12 (cum-logit)					
Intercept 1: Yes	-1.84755	0.330484	1.00	-5.59	0.0000
DARE Program					
Yes	-0.52248	0.182076	1.00	-2.87	0.0069
No	0.000000	0.000000	.	.	.
Grade in School					
5th Grade	-0.50020	0.223481	1.00	-2.24	0.0317
6th Grade	0.000000	0.000000	.	.	.
SEX					
Male	0.084014	0.181161	1.00	0.46	0.6457
Female	0.000000	0.000000	.	.	.
RACE					
Black	0.497135	0.283813	1.00	1.75	0.0886
Hispanic	0.095132	0.386261	1.00	0.25	0.8069
Other	0.493601	0.282336	1.00	1.75	0.0892
White	0.000000	0.000000	.	.	.
Family Situation					
Non-Traditional	0.420841	0.192760	1.00	2.18	0.0358
Traditional	0.000000	0.000000	.	.	.
AREA					
Rural	-0.07878	0.319795	1.00	-0.25	0.8069
Suburban	-0.25078	0.271144	1.00	-0.92	0.3613
Urban	0.000000	0.000000	.	.	.

Here we see that 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 not valid for the data at hand. They merely demonstrate the consequences of ignoring the experimental design. The design effects are all equal to 1.0, since both numerator and denominator values are the same.

Results

GEE with Independent "Working" Correlations
Model-Based (Naive) Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 2
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation Method: Model-Based (Naive)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Model-Based Variance Estimation
```

Contrast	Degrees of Freedom	Wald ChiSq	P-value Wald ChiSq
OVERALL MODEL	10	470.28	0.0000
MODEL MINUS INTERCEPT	9	30.41	0.0004
DARE	1	8.23	0.0041
FIFTH	1	5.01	0.0252
SEX	1	0.22	0.6428
RACE	3	4.64	0.2001
OTHFAM	1	4.77	0.0290
AREA	2	1.02	0.6011

This table contains the main effects tests computed as if the naive assumption of independence were true. The Wald chi-square test is used to evaluate the null hypotheses.

Results

GEE with Independent "Working" Correlations
Model-Based (Naive) Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 3
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation Method: Model-Based (Naive)
Working Correlations: Independent
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

Model-Based Variance Estimation

------------------------------------------------
Independent Variables                Lower   Upper
  and Effects              Odds   95%     95%
                           Ratio  Limit   Limit
------------------------------------------------
INTCIG12 (cum-logit)
  Intercept 1: Yes          0.16    0.08    0.31
DARE Program
  Yes                       0.59    0.41    0.86
  No                        1.00    1.00    1.00
Grade in School
  5th Grade                 0.61    0.39    0.95
  6th Grade                 1.00    1.00    1.00
SEX
  Male                      1.09    0.75    1.57
  Female                    1.00    1.00    1.00
RACE
  Black                     1.64    0.92    2.93
  Hispanic                  1.10    0.50    2.41
  Other                     1.64    0.92    2.91
  White                     1.00    1.00    1.00
Family Situation
  Non-Traditional           1.52    1.03    2.25
  Traditional               1.00    1.00    1.00
AREA
  Rural                     0.92    0.48    1.77
  Suburban                  0.78    0.45    1.35
  Urban                     1.00    1.00    1.00
------------------------------------------------

MULTILOG used
  CPU time       : 6.81 seconds
  Elapsed time   : 7 seconds
  Virtual memory : 1.52 MB
```

Results

GEE with Exchangeable "Working" Correlations
Robust Variance Estimator

This next set of SUDAAN programming statements fits 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. All other programming statements remain the same as previously.

```
32  PROC MULTILOG DATA="c:\\tera\\examples\\DARE" FILETYPE=SAS
                  SEMETHOD=ZEGER R=EXCHANGE;

33  NEST _ONE_ SCHOOL;

34  WEIGHT _ONE_;

35  SUBGROUP DARE FIFTH SEX RACE OTHFAM AREA INTCIG12;

36  LEVELS   2    2     2   4    2      3    2;

37  MODEL INTCIG12 = DARE FIFTH SEX RACE OTHFAM AREA / CUMLOGIT;

38  TEST WALDCHI;

39  SETENV LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60 COLSPCE=2;

40  PRINT  BETA="BETA" SEBETA="STDERR" T_BETA="T:BETA=0" P_BETA="P-VALUE"
           / RISK=ALL TESTS=DEFAULT RHOS=ALL BETAFMT=F8.6 SEBETAFMT=F8.6
             T_BETAFMT=F8.2 WALDCHIFMT=F6.2 WALDCHPFMT=F7.4 DFFMT=F7.0
             ORFMT=F5.2 LOWORFMT=F6.2 UPORFMT=F6.2;

41  TITLE "MULTILOG Logistic Regression Model for the DARE Evaluation Study"
          "Ennett, et al, 1994";

Opened SAS data file c:\tera\examples\DARE.SSD for reading.

Number of observations read       :   1525    Weighted count:     1525
Observations used in the analysis :   1188    Weighted count:     1188
Observations with missing values  :    337    Weighted count:      337
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is 10
File c:\tera\examples\DARE.SSD contains   36 Clusters
Maximum cluster size is 153 records
Minimum cluster size is  11 records

Independence parameters have converged in 4 iterations
Step 1 parameters have converged in 6 iterations.

Sample and Population Counts for Response Variable INTCIG12
  Yes:  Sample Count      145    Population Count       145
  No :  Sample Count     1043    Population Count      1043
```

By default, SUDAAN fits the 1-step GEE estimates (Lipsitz, et al., 1994). Here we see that the independence betas (the starting estimates for GEE exchangeable) have converged in 4 iterations, and the Step 1 GEE parameter estimates (under exchangeable working correlations) have converged in 6 iterations.

Results

GEE with Exchangeable "Working" Correlations
Robust Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 1
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation Method: Robust (Zeger-Liang, 1986)
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    Initiation of Cigarette Use
  Use                              Yes
------------------------------------------------------
Yes                          0.0206
------------------------------------------------------
```

This table contains 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.

Note that although the intracluster correlation is small, the cluster sizes in this study are large enough to cause almost a doubling in the variance of estimated regression coefficients (deff=1.75 for the DARE effect in the working independence model with robust variance estimate).

Results

GEE with Exchangeable "Working" Correlations
Robust Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 2
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation Method: Robust (Zeger-Liang, 1986)
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    STDERR   T:BETA=0   P-VALUE
------------------------------------------------------------------
INTCIG12 (cum-logit)
  Intercept 1: Yes         -1.88017  0.449771      -4.18    0.0002
DARE Program
  Yes                      -0.58250  0.242184      -2.41    0.0216
  No                       0.000000  0.000000        .        .
Grade in School
  5th Grade                -0.46289  0.221616      -2.09    0.0441
  6th Grade                0.000000  0.000000        .        .
SEX
  Male                     0.087569  0.157590       0.56    0.5820
  Female                   0.000000  0.000000        .        .
RACE
  Black                    0.508801  0.367707       1.38    0.1752
  Hispanic                 0.277801  0.412405       0.67    0.5050
  Other                    0.518041  0.427964       1.21    0.2342
  White                    0.000000  0.000000        .        .
Family Situation
  Non-Traditional          0.436618  0.173405       2.52    0.0165
  Traditional              0.000000  0.000000        .        .
AREA
  Rural                    -0.06764  0.378772      -0.18    0.8593
  Suburban                 -0.26165  0.371397      -0.70    0.4858
  Urban                    0.000000  0.000000        .        .
------------------------------------------------------------------
```

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 independence (-0.5225). The variance estimate has also increased, but only slightly. Nevertheless, the overall conclusions are qualitatively the same as for independent working correlations with a robust variance estimate.

Results

GEE with Exchangeable "Working" Correlations
Robust Variance Estimator

```
Date: 04-28-97            Research Triangle Institute             Page  : 3
Time: 13:30:35              The MULTILOG Procedure                Table : 1

Variance Estimation Method: Robust (Zeger-Liang, 1986)
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         P-value
                           of       Wald    Wald
                           Freedom  ChiSq   ChiSq
---------------------------------------------------
OVERALL MODEL                   10  312.35   0.0000
MODEL MINUS INTERCEPT            9   29.80   0.0005
DARE                             1    5.78   0.0162
FIFTH                            1    4.36   0.0367
SEX                              1    0.31   0.5784
RACE                             3    1.93   0.5867
OTHFAM                           1    6.34   0.0118
AREA                             2    0.58   0.7497
---------------------------------------------------
```

Here we see the model main effects, under the exchangeable option and a robust variance estimate. All of the effects have become slightly more significant compared to working independence with a robust variance estimate. This should not be taken as a general result for exchangeability vs. working independence. Studies have shown that modelling the correlations tend to yield greater power for detecting within-cluster covariates (Neuhaus, 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 modelling the correlation structure.

Results

GEE with Exchangeable "Working" Correlations
Robust Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 4
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation Method: Robust (Zeger-Liang, 1986)
Working Correlations: Exchangeable
Link Function: Cumulative Logit
Response variable INTCIG12: Initiation of Cigarette Use

MULTILOG Logistic Regression Model for the DARE Evaluation Study

------------------------------------------------
Independent Variables              Lower   Upper
  and Effects             Odds      95%     95%
                          Ratio   Limit   Limit
------------------------------------------------
INTCIG12 (cum-logit)
  Intercept 1: Yes         0.15    0.06    0.38
DARE Program
  Yes                      0.56    0.34    0.91
  No                       1.00    1.00    1.00
Grade in School
  5th Grade                0.63    0.40    0.99
  6th Grade                1.00    1.00    1.00
SEX
  Male                     1.09    0.79    1.50
  Female                   1.00    1.00    1.00
RACE
  Black                    1.66    0.79    3.51
  Hispanic                 1.32    0.57    3.05
  Other                    1.68    0.70    4.00
  White                    1.00    1.00    1.00
Family Situation
  Non-Traditional          1.55    1.09    2.20
  Traditional              1.00    1.00    1.00
AREA
  Rural                    0.93    0.43    2.02
  Suburban                 0.77    0.36    1.64
  Urban                    1.00    1.00    1.00
------------------------------------------------

MULTILOG used
  CPU time       : 30.15 seconds
  Elapsed time   : 31 seconds
  Virtual memory : 1.58 MB
```

The estimated odds of initiating smoking by Wave 2 is now 0.56 under exchangeability, vs. 0.59 under working independence. It should also be noted that for binary outcomes, GEE results under the GENLOGIT and CUMLOGIT options are identical, since they specify the same model.

Results

GEE with Exchangeable "Working" Correlations Model-Based (Naive) Variance Estimator

Below are results from the exchangeable correlation model using the ***model-based*** or ***naive variance-covariance matrix*** of the estimated regression coefficients. The model-based variance is the M_0^{-1} matrix, or the outside portion of the robust variance estimate: $M_0^{-1} = [\boldsymbol{D}'\boldsymbol{V}^{-1}\boldsymbol{D}]^{-1}$, 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 ***assuming that the exchangeable "working" correlation assumption were correct***. Since that is close to truth for students clustered within schools, we will see that results are essentially the same as with the robust variance estimator.

```
42  PROC MULTILOG DATA="c:\\tera\\examples\\DARE" FILETYPE=SAS
                  SEMETHOD=MODEL R=EXCHANGE;

43  NEST _ONE_ SCHOOL;

44  WEIGHT _ONE_;

45  SUBGROUP DARE FIFTH SEX RACE OTHFAM AREA INTCIG12;

46  LEVELS   2    2     2   4    2      3    2;

47  MODEL INTCIG12 = DARE FIFTH SEX RACE OTHFAM AREA / CUMLOGIT;

48  TEST WALDCHI;

49  SETENV LABWIDTH=25 COLWIDTH=8 DECWIDTH=4 LINESIZE=78 PAGESIZE=60 COLSPCE=2;

50  PRINT  BETA="BETA" SEBETA="STDERR" T_BETA="T:BETA=0"
           P_BETA="P-VALUE" / RISK=ALL TESTS=DEFAULT RHOS=ALL
           BETAFMT=F8.6 SEBETAFMT=F8.6 T_BETAFMT=F8.2 WALDCHIFMT=F6.2
           WALDCHPFMT=F7.4 DFFMT=F7.0
           ORFMT=F5.2 LOWORFMT=F6.2 UPORFMT=F6.2;

51  TITLE "MULTILOG Logistic Regression Model for the DARE Evaluation Study"
          "Model-Based Variance Estimation";

Opened SAS data file c:\tera\examples\DARE.SSD for reading.

Number of observations read       :   1525    Weighted count:     1525
Observations used in the analysis :   1188    Weighted count:     1188
Observations with missing values  :    337    Weighted count:      337
Denominator degrees of freedom    :     35

Maximum number of estimable parameters for the model is 10

File c:\tera\examples\DARE.SSD contains   36 Clusters
Maximum cluster size is 153 records
```

```
Minimum cluster size is  11 records

Independence parameters have converged in 4 iterations

Step 1 parameters have converged in 6 iterations.

Sample and Population Counts for Response Variable INTCIG12
  Yes:  Sample Count      145    Population Count       145
  No :  Sample Count     1043    Population Count      1043
```

Results

GEE with Exchangeable "Working" Correlations
Model-Based (Naive) Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 1
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation 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

Model-Based Variance Estimation

-----------------------------------------------------------------
Independent Variables
  and Effects                   BETA    STDERR  T:BETA=0   P-VALUE
-----------------------------------------------------------------
INTCIG12 (cum-logit)
  Intercept 1: Yes          -1.88017  0.369879     -5.08    0.0000
DARE Program
  Yes                       -0.58250  0.243284     -2.39    0.0221
  No                        0.000000  0.000000       .        .
Grade in School
  5th Grade                 -0.46289  0.270334     -1.71    0.0957
  6th Grade                 0.000000  0.000000       .        .
SEX
  Male                      0.087569  0.181267      0.48    0.6320
  Female                    0.000000  0.000000       .        .
RACE
  Black                     0.508801  0.304171      1.67    0.1033
  Hispanic                  0.277801  0.379479      0.73    0.4690
  Other                     0.518041  0.287578      1.80    0.0803
  White                     0.000000  0.000000       .        .
Family Situation
  Non-Traditional           0.436618  0.193741      2.25    0.0306
  Traditional               0.000000  0.000000       .        .
AREA
  Rural                     -0.06764  0.375932     -0.18    0.8582
  Suburban                  -0.26165  0.344689     -0.76    0.4529
  Urban                     0.000000  0.000000       .        .
-----------------------------------------------------------------
```

Here we have 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, indicating that the exchangeable correlation assumption is close to truth.

Results

GEE with Exchangeable "Working" Correlations
Model-Based (Naive) Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 2
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation 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

Model-Based Variance Estimation

---------------------------------------------------
Contrast                   Degrees          P-value
                           of       Wald    Wald
                           Freedom  ChiSq   ChiSq
---------------------------------------------------
OVERALL MODEL                   10  262.19   0.0000
MODEL MINUS INTERCEPT            9   22.36   0.0078
DARE                             1    5.73   0.0167
FIFTH                            1    2.93   0.0868
SEX                              1    0.23   0.6290
RACE                             3    4.06   0.2546
OTHFAM                           1    5.08   0.0242
AREA                             2    0.64   0.7249
---------------------------------------------------
```

Here we have the ***main effects tests*** computed under exchangeability, using the model-based variance approach. Results are essentially the same as with the robust variance estimator.

Results

GEE with Exchangeable "Working" Correlations
Model-Based (Naive) Variance Estimator

```
Date: 04-28-97            Research Triangle Institute            Page  : 3
Time: 13:30:35              The MULTILOG Procedure               Table : 1

Variance Estimation 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

Model-Based Variance Estimation

------------------------------------------------
Independent Variables                Lower   Upper
  and Effects                Odds     95%     95%
                             Ratio   Limit   Limit
------------------------------------------------
INTCIG12 (cum-logit)
  Intercept 1: Yes            0.15    0.07    0.32
DARE Program
  Yes                         0.56    0.34    0.92
  No                          1.00    1.00    1.00
Grade in School
  5th Grade                   0.63    0.36    1.09
  6th Grade                   1.00    1.00    1.00
SEX
  Male                        1.09    0.76    1.58
  Female                      1.00    1.00    1.00
RACE
  Black                       1.66    0.90    3.08
  Hispanic                    1.32    0.61    2.85
  Other                       1.68    0.94    3.01
  White                       1.00    1.00    1.00
Family Situation
  Non-Traditional             1.55    1.04    2.29
  Traditional                 1.00    1.00    1.00
AREA
  Rural                       0.93    0.44    2.01
  Suburban                    0.77    0.38    1.55
  Urban                       1.00    1.00    1.00
------------------------------------------------
```

References

Applications of SUDAAN and Related Techniques

Bieler, G. and Williams, R. (1995). Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics* **51**, 764-776.

Davies, G.M. (1994). Applications of sample survey methodology to repeated measures data structures in dentistry. *Institute of Statistics Mimeo Series* No. 2128T. University of North Carolina: Chapel Hill.

Donner, A. (1982). An empirical study of cluster randomization. *American Journal of Epidemiology* **11**, 283-286.

Ennett, S.T., Rosenbaum, D.P., Flewelling, R.L., Bieler, G.S., Ringwalt, C.L., and Bailey, S.L. (1994). Long-term evaluation of drug abuse resistance education. *Addictive Behaviors*, **19**, 113-125.

Fung, K.Y., Krewski, D., and Scott, A.J. (1994). Tests for trend in developmental toxicity experiments with correlated binary data. *Risk Analysis* **14**, 639-648.

Gansky, SA, Koch, GG, and Wilson, J. (1994). Statistical evaluation of relationships between analgesic dose and ordered ratings of pain relief over an eight-hour period. *Journal of Biopharmaceutical Statistics* **4**, 233-265.

Graubard, B.I. and Korn, E.L. (1994). Regression analysis with clustered data. *Statistics in Medicine* **13**, 509-522.

LaVange, L.M. and Koch, G.G. (1994). Analysis of repeated measures studies with multiple regression methods for sample survey data. Presented and the 1994 Drug Information Association Meetings, and submitted for publication.

LaVange, L.M., Keyes, L.L., Koch, G.G., and Margolis, P.A. (1994). Application of sample survey methods for modelling ratios to incidence densities. *Statistics in Medicine* **13**, 343-355.

Norton, E., Bieler, G., Zarkin, G., and Ennett, S. (1996). Analysis of prevention program effectiveness with clustered data using generalized estimating equations. *Journal of Consulting and Clinical Psychology* **64**, 919-926.

Rao, J. and Colin, D. (1991). Fitting dose-response models and hypothesis testing in teratological studies. In: <u>Statistics in Toxicology</u>, Krewski and Franklin, eds. NY: Gordon and Breach.

Rao, J. and Scott, A.J. (1992). A simple method for the analysis of clustered binary data. *Biometrics* **48**, 577-585.

Schmid, JE, Koch, GG, and LaVange, LE (1991). An overview of statistical issues and methods of meta-analysis. *Journal of Biopharmaceutical Statistics*, **1**, 103-120.

Williams, R.L. (1995). Product-limit survival functions with correlated survival times. *Lifetime Data Analysis* **1**, 171-186.

Williams, R.L. and Bieler, G.S. (1993). Estimation of proportional hazards models for survival times with nested errors. ASA Proceedings of the Biopharmaceutical Section, 115-120.

Survey Sampling

Binder, D. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* **79**, 139-147.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279-292.

Cochran, W.G. (1977). Sampling Techniques. Wiley, New York.

Folsom, R.E. (1974). National Assessment Approach to Sampling Error Estimation, Sampling Error Monograph. Prepared for the National Assessment of Educational Progress, Denver, CO.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Volume I: Methods and Applications. NY: Wiley.

Kendall, M.G. and Stuart, A. (1973). The Advanced Theory of Statistics. NY: Hafner Publishing Co.

Kish, L. (1965). Survey Sampling. NY: Wiley.

Kish, L. and Frankel, M. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society*, Series B, **36**, 1-37.

Koch, G.G., Freeman, D.H. and Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. *International Statistical Review* **43**, 59-78.

LaVange, L.M., Iannacchione, V.G., and Garfinkel, S. (1986). An application of logistic regression methods to survey data: predicting high cost users of medical care. American Statistical Association, Proceedings of the Section on Survey Research Methods, Washington, DC.

Rao, J. And Scott, A. (1987). On simple adjustments to chi-squared tests with sample survey data. Annals of Statistics, **15**, 385-397.

Sarndal, C., Swensson, B., and Wretman, J. (1992). Model-Assisted Survey Sampling. Springer-Verlag, NY.

Scott, A.J. and Holt, D. (1982). The effect of 2-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77**, 848-854.

Shah, B.V., Barnwell, B.G., and Bieler, G.S. (1996). SUDAAN User's Manual, Release 7, First Edition. Research Triangle Institute, RTP, NC.

Shah, B.V. and LaVange, L.M. (1994). Mixed models for survey data. American Statistical Association,
Proceedings of the Section on Survey Research Methods.

Shah, B.V., Holt, M.M. and Folsom, R.E. (1977). Inference about regression models from sample survey
data. Bulletin of the International Statistical Institute XLVII, **3**, 43-57.

Thomas, D.R. and Rao, J.N.K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *JASA* **82**, 630-636.

Wolter, K.M. (1985). Introduction to Variance Estimation. NY: Springer-Verlag.

Woodruff, R. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* **66**, 411-414.

GEEs and Generalized Linear Models

Carr, G. and Portier, C. (1993). An evaluation of some methods for fitting dose-response models to quantal response data. *Biometrics* **49**, 779-791.

Diggle, P., Liang, K.Y., and Zeger, S.L. (1994). Analysis of Longitudinal Data. NY: Oxford University Press.

Dunlop, D.D. (1994). Regression for longitudinal data: a bridge from least squares regression. *American Statistician* **48**, 299-303.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability **1**, 221-233.

Karim, M.R. and Zeger, S.L. (1989). GEE: A SAS macro for longitudinal data analysis. Technical Report #674 from the Department of Biostatistics, The Johns Hopkins University.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

Lipsitz, S.R., Kim, K. and Zhao, L. (1994a). Analysis of repeated categorical data using generalized
stimating equations. *Statistics in Medicine* **13**, 1149-1163.

Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J. and Laird, N.M. (1994b). Performance of generalized estimating equations in practical situations. *Biometrics* **50**, 270-278.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. NY: Chapman and Hall.

Neuhaus, J.M. and Segal, M.R. (1993). Design effects for binary regression models fitted to dependent
data. *Statistics in Medicine* **12**, 1259-1268.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485-497.

Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review* **54**, 221-226.

Zeger, S. (1988). Commentary. *Statistics in Medicine* **7**, 161-168.

Zeger, S. and Liang, K. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Jackknife Variance Estimators

Carr, G. And Portier, C. (1993). An evaluation of some methods for fitting dose-response models to quantal-response developmental toxicity data. *Biometrics* **49**, 779-791.

Lipsitz, S. And Parzen, M. (1996). A jackknife estimator of variance for Cox regression for correlated survival data. *Biometrics* **52**, 291-298.

Lipsitz, S., Dear, K., and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics* **50**, 842-846.

Quenouille, M. (1956). Notes on bias in estimation. *Biometrika* **43**, 353-360.

Tukey, J.W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, **29**, 614.

BRR Variance Estimators

McCarthy, PJ (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. *Vital and Health Statistics*, Series 2, No. 14, National Center for Health Statistics, Public Health Service, Washington, DC.

McCarthy, PJ (1969). Pseudoreplication: Half-samples. *Review of the International Statistical Institute* **37**, 239-264.

Plackett, RL and Burman, JP (1946). The Design of Optimum Multifactorial Experiments. *Biometrika* **33**, 305-325.

Wolter, KM (1985). Introduction to Variance Estimation. NY: Springer-Verlag.