Additive interaction in SUDAAN

Prepared by Aaron Sarvet and Melanie Wall at Columbia University, Research Foundation for Mental Hygiene, and New York State Psychiatric Institute on 08/03/2016

It is becoming more common for investigators to investigate interaction on the additive scale for binary outcomes. With binary outcomes, the additive scale means that comparisons are made in terms of *risk differences* rather than risk ratios or odds ratios. Hence, an interaction on the additive scale is present when the risk difference for some predictor (exposure variable) on an outcome varies across some other predictor variable (the effect modifier). This manual illustrates how to test interaction on the additive scale in SUDAAN, a SAS-callable statistical package commonly used for the analysis of complex survey data. The manual will be split into two sections: 1) testing additive interaction with two categorical predictors; and 2) testing additive interaction with a one categorical predictor and one continuous predictor. Estimation of the model adjusted risks and adjusted risk differences as well as the test of additive interaction (i.e. difference in risk differences) is done using the predictive margins functions PREDMARG and PRED_EFF in SUDAAN within PROC RLOGIST. Specifically, predicted probabilities from the logistic model including covariates are obtained for each risk group comparison while allowing all the other covariates to vary according to their observed values for each person and then averaged (weighted averaged if there are sampling weights). Note, this predictive margins approach is different than the conditional margins approach (CONDMARG and COND_EFF) where all the covariates are fixed at their mean value. For more information on predicted marginal prevalences, see Graubard and Korn (1999) or Bieler et al. (2010) or SUDAAN manual for RLOGIST.

<u>Section 1: testing additive interaction with two categorical predictors:</u>

The following example will utilize NSDUH (National Survey on Drug Use and Health) data (N= 492831) from 2002-2014. The example will use 7 analytic variables and 3 complex survey variables.

Analytic variables

- MRJYR: Whether or not the respondent had used marijuana in the past year (1=yes; 0=no) – *Outcome*
- RDIFMJ: Perception of the respondent that marijuana is easy or fairly easy to obtain (1=yes; 0=no) – *Primary exposure variable*
- IRSEX: The sex of the respondent (1=male; 2=female) – *Potential effect modifier*
- YEAR: Calendar year (continuous, 2002-2014) – *Control variable*
- INCOME: Annual income of the respondent (1=<$10k; 2=$10-20k; 3=$20-40k; 4=$40+k) – *Control variable*
- EDUCCAT2: Educational attainment of the respondent (1=<HS; 2=HS; 3>HS) – *Control variable*

Complex survey variables

- VESTR: sample stratum
- VEREP: sample PSU
- Analwt_new: sample weight

Research question: How does the association between ease of obtainment (RDIFMJ) and marijuana use (MRJYR) vary by sex (IRSEX) on the *additive scale*, after controlling for year, income and education?

We fit the following model:

```
proc rlogist DESIGN=WR DATA=data; NEST VESTR VEREP / MISSUNIT; WEIGHT
analwt_new;
class RDIFMJ IRSEX INCOME EDUCCAT2;
model MRJYR =  RDIFMJ IRSEX RDIFMJ*IRSEX YEAR INCOME EDUCCAT2;

predmarg RDIFMJ*IRSEX;

pred_eff RDIFMJ=(-1 1)*IRSEX=(1 0) / name ="Males: Difference in prevalence of
marijauana use between those who easily and can't easily obtain marijuana" ;

pred_eff RDIFMJ=(-1 1)*IRSEX=(0 1) / name ="Females: Difference in prevalence
of marijauana use between those who easily and can't easily obtain marijuana" ;

pred_eff RDIFMJ=(-1 1)*IRSEX=(1 -1) / name ="Test of additive interaction:
difference-in-difference" ;

run;
```

IRSEX and RDIFMJ are on the *class* statement, indicating that they are analyzed as categorical variables.
The *predmarg* statement calls for the computation and display of predicted marginal prevalences of marijuana use within strata of sex and ease of obtainment. See below for the predicted marginal prevalences output by SUDAAN:

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable MRJYR: MARIJUANA - PAST YEAR USE
by: Predicted Marginal #1.

Predicted Marginal #1	Predicted Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0
MARIJUANA FAIRLY OR EASY TO OBTAIN, IMPUTATION REVISED GENDER					
0, 1	0.04498	0.00092	0.04320	0.04682	48.99251
0, 2	0.01929	0.00058	0.01816	0.02047	32.97182
1, 1	0.21328	0.00159	0.21015	0.21644	134.01175
1, 2	0.13724	0.00130	0.13469	0.13983	105.39112

So for men who do not think MJ is easy to obtain (0,1) the prevalence of MJ use is 4.498% and for women who do not think MJ is easy to obtain (0,2) the prevalence of MJ use is 1.929%. Whereas among those who do think MJ is easy to obtain, men (1,1) use at 21.328% and women (1,2) 13.724%.

The *pred_eff* statement calls for the contrasts of these predicted marginal prevalences. The first *pred_eff* statement calls for a contrast between the first and second level of RDIFMJ (easy vs. not easy to obtain) within the first level of sex (males). That is, it calls for a test of the difference between 0.21328 and 0.04498 which is 0.16830. Here is the output below:

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable MRJYR: MARIJUANA - PAST YEAR USE
by: Contrasted Predicted Marginal #1.

-------------------------------------------------------------------
Contrasted
  Predicted
  Marginal     PREDMARG
  #1           Contrast             SE       T-Stat      P-value
-------------------------------------------------------------------
Males:
  Difference
   in
  prevalence
  of
  marijauana
  use
  betweeen
  those who
  easily and
  can't
  easily
  obtain
  marijuana       0.16830      0.00178     94.34479      0.00000
-------------------------------------------------------------------
```

The second *pred_eff* statement calls for a contrast between the first and second level of RDIFMJ (easy vs. not easy to obtain) within the second level of sex (females). That is, it calls for a test of the difference between 0.13724 and 0.01929 which is 0.11795. Here is the output below:

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable MRJYR: MARIJUANA - PAST YEAR USE
by: Contrasted Predicted Marginal #2.

-------------------------------------------------------------------
Contrasted
  Predicted
  Marginal     PREDMARG
  #2           Contrast             SE       T-Stat      P-value
-------------------------------------------------------------------
Females:
  Difference
   in
  prevalence
  of
  marijauana
  use
  betweeen
  those who
  easily and
  can't
  easily
  obtain
  marijuana       0.11795      0.00141     83.73545      0.00000
```

The third *pred_eff* statement calls for a contrast of two differences: the difference between the first and second level of RDIFMJ (easy vs. not easy to obtain) within the first level of sex (males); and the difference between the first and second level of RDIFMJ (easy vs. not easy to obtain) within the second level of sex (females). That is, it calls for a test of the difference between (0.21328 - 0.04498 – i.e. the first pred_eff computation) and (0.13724 - 0.01929 – i.e. the second pred_eff computation). Specifically, the comparison is .16830 - .11795 = .05035. This is the test for interaction on the additive scale – the difference-in-difference. Here is the output below:

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable MRJYR: MARIJUANA - PAST YEAR USE
by: Contrasted Predicted Marginal #3.

----------------------------------------------------------------
Contrasted
  Predicted
  Marginal     PREDMARG
  #3           Contrast              SE        T-Stat       P-value
----------------------------------------------------------------
Test of
  additive
  interacti-
  on:
  difference
  -in-
  difference       0.05035      0.00209      24.13360       0.00000
----------------------------------------------------------------
```

The p-value for the additive interaction is significant indicating that there is a larger effect in men than women of easily available marijuana has on them using it.

Section 2: testing additive interaction with one categorical predictor and one continuous predictor:

The following example will again utilize NSDUH (National Survey on Drug Use and Health) data (N= 492831) from 2002-2014 and similar variables to the previous section.

Research question: How do trends in marijuana use (MRJYR) over time (YEAR, a continuous variable) vary by sex (IRSEX) on the *additive scale*, after controlling for income and education?

We fit the following model:

```
proc rlogist DESIGN=WR DATA=data; NEST VESTR VEREP / MISSUNIT; WEIGHT analwt_new;
class IRSEX INCOME EDUCCAT2;
model MRJYR = IRSEX IRSEX*YEAR INCOME EDUCCAT2;

predmarg YEAR*IRSEX / YEAR=(2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010,
2011, 2012, 2013, 2014);

pred_eff YEAR=(-1 1)*IRSEX=(1 0) / year=(2002, 2014) name ="Males: Difference in
prevalence of marijauana use between 2002 and 2014" ;

pred_eff YEAR=(-1 1)*IRSEX=(0 1) / year=(2002, 2014) name ="Females: Difference in
prevalence of marijauana use between 2002 and 2014" ;
```

```
pred_eff YEAR=(-1 1)*IRSEX=(1 -1) / year=(2002, 2014) name ="Test of additive
interaction: difference-in-difference" ;

run;
```

IRSEX is on the *class* statement, indicating that it is analyzed as a categorical variable. However, YEAR is not on the class statement, indicating that it is analyzed as continuous linear predictor of the log-odds of marijuana use. As before, the *predmarg* statement calls for the computation and display of predicted marginal prevalences of marijuana use within strata of sex and year. However, since YEAR is a continuous variable, the user has to specify which values for YEAR to compute predicted marginal for. In this example, all years from 2002-2014 are requested. Here is the output below:

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable MRJYR: MARIJUANA - PAST YEAR USE
by: Predicted Marginal #1.
```

Predicted Marginal #1	Predicted Marginal	SE	Lower 95% Limit	Upper 95% Limit	T:Marg=0
IMPUTATION REVISED GENDER, YEAR					
1, 2002	0.12460	0.00181	0.12106	0.12821	68.79141
1, 2003	0.12750	0.00164	0.12430	0.13077	77.76733
1, 2004	0.13046	0.00147	0.12758	0.13339	88.67464
1, 2005	0.13347	0.00131	0.13090	0.13608	101.72016
1, 2006	0.13655	0.00117	0.13425	0.13888	116.54811
1, 2007	0.13968	0.00106	0.13759	0.14179	131.37321
1, 2008	0.14287	0.00100	0.14090	0.14487	142.19556
1, 2009	0.14612	0.00101	0.14414	0.14813	144.32782
1, 2010	0.14944	0.00109	0.14729	0.15161	136.81372
1, 2011	0.15281	0.00124	0.15039	0.15527	123.52106
1, 2012	0.15625	0.00143	0.15344	0.15910	108.98945
1, 2013	0.15975	0.00167	0.15648	0.16307	95.70630
1, 2014	0.16331	0.00193	0.15952	0.16716	84.41656
2, 2002	0.06798	0.00105	0.06593	0.07008	64.61380
2, 2003	0.06980	0.00096	0.06794	0.07171	73.04862
2, 2004	0.07167	0.00087	0.06998	0.07339	82.83241
2, 2005	0.07358	0.00079	0.07204	0.07515	93.47884
2, 2006	0.07554	0.00073	0.07411	0.07699	103.49774
2, 2007	0.07754	0.00070	0.07617	0.07894	110.18517
2, 2008	0.07960	0.00072	0.07819	0.08103	110.87854
2, 2009	0.08170	0.00078	0.08018	0.08325	105.33081
2, 2010	0.08386	0.00087	0.08215	0.08560	95.95763
2, 2011	0.08606	0.00101	0.08410	0.08807	85.56260
2, 2012	0.08832	0.00116	0.08605	0.09065	75.82071
2, 2013	0.09063	0.00135	0.08801	0.09332	67.34261
2, 2014	0.09300	0.00155	0.08999	0.09609	60.18298

Again, the *pred_eff* statement can call for the contrasts of these predicted marginal prevalences. When a continuous variable is used in a *pred_eff* statement, the user must specify which levels of the continuous variable are to be used in the contrast. In this example, the two endpoint years (2002, 2014) are specified. The first *pred_eff* statement calls for a contrast between the first and second specified level of YEAR (2002 vs. 2014) within the first level of sex (males). That is, it calls for a test of the difference between 0.16331 and 0.12460 which is 0.03871. Here is the output below:

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable MRJYR: MARIJUANA - PAST YEAR USE
by: Contrasted Predicted Marginal #1.

----------------------------------------------------------------
Contrasted
  Predicted
  Marginal     PREDMARG
  #1           Contrast              SE       T-Stat      P-value
----------------------------------------------------------------
Males:
  Difference
   in
  prevalence
  of
  marijauana
  use
  between
  2002 and
  2014            0.03871      0.00318     12.17820      0.00000
----------------------------------------------------------------
```

The second *pred_eff* statement calls for a contrast between the first and second specified level of YEAR (2002 vs. 2014) within the second level of sex (females). That is, it calls for a test of the difference between 0.09300 and 0.06798 which is .02502. Here is the output below:

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable MRJYR: MARIJUANA - PAST YEAR USE
by: Contrasted Predicted Marginal #2.

----------------------------------------------------------------
Contrasted
  Predicted
  Marginal     PREDMARG
  #2           Contrast              SE       T-Stat      P-value
----------------------------------------------------------------
Females:
  Difference
   in
  prevalence
  of
  marijauana
  use
  between
  2002 and
  2014            0.02502      0.00217     11.52920      0.00000
```

The third *pred_eff* statement calls for a contrast of two differences: the difference between the first and second specified level of YEAR (2002 vs. 2014) within the first level of sex (males); and the difference between the first and second specified level of YEAR (2002 vs. 2014) within the second level of sex (females). That is, it calls for a test of the difference between (0.16331- 0.12460– i.e. the first pred_eff computation) and (0.09300 - 0.06798– i.e. the second pred_eff computation). This is the test for differential trends from 2002 to 2014 in use between men and women on the additive scale (i.e. the additive interaction test or the difference-in-difference). Here is the output below:

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable MRJYR: MARIJUANA - PAST YEAR USE
by: Contrasted Predicted Marginal #3.

---------------------------------------------------------------
Contrasted
  Predicted
  Marginal     PREDMARG
  #3           Contrast            SE       T-Stat      P-value
---------------------------------------------------------------
Test of
  additive
  interacti-
  on:
  difference
  -in-
  difference      0.01369     0.00358     3.82198      0.00019
---------------------------------------------------------------
```

So we find the additive interaction to be statistically significant indicating that the increase in prevalence of MJ use in Men over the time period is greater than the increase in Females.

References:

- Graubard B, Korn E (1999) "Predictive Margins with Survey Data" Biometrics 55:652-659
- Bieler, Brown, Williams, & Brogan (2010) "Estimating Model-Adjusted Risks, Risk Differences, and Risk Ratios From Complex Survey Data" Am J Epi DOI: 10.1093/aje/kwp440, here.
- RLOGIST example #3 at http://sudaansupport.rti.org/page.cfm/SUDAAN_RLOGIST