

Multivariate Log-binomial (and Poisson) Regression between Years since Quitting Smoking and Self-reported Oral Health:

A Sample of Work in SAS

Methods

This study used data from the Canadian Community Health Survey (CCHS), Annual Component, from 2015-2016. A subset was taken of the CCHS, consisting of 13 variables:

- Sex
- Age
- Education
- BMI (self-reported)
- Diabetes
- Number of years since stopped smoking completely (**← the exposure**)
 - o Levels: 0 (<1 year), 2 (1 to 2 years), 3 (3 to 5 years), 4 (>=11 years)
- Self-perceived oral health (**← the outcome**)
 - o Levels: 1 (poor), 2 (fair), 3 (good), 4 (very good), 5 (excellent)
- Frequency of tooth brushing
- Frequency of dental visits
- Diet
- Alcohol
- Drug use
- Income

The sample contained 109,659 observations in total (including missing values).

Data Analysis:

```
**** Loading the dataset, renaming the variables. ;
LIBNAME smoking "C:/Users/namakuto/Documents/Data Management/Assignment 4";
/*//////////////////// .csv IMPORT and SAS file MERGING //////////////////*/
FILENAME REFFILE "C:/Users/namakuto/Documents/Data Management/Assignment
4/cchs.csv";
PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV OUT=smoking.data1; GETNAMES=YES;RUN;
data smoking.data1; set smoking.data1;
rename ALC_015=Alcohol CCC_095=Diabetes DEN_030=DentalVisits
DEN_010A=Toothbrushing DHHGAGE=Age DHH_SEX=Sex DRMDVLAY=Drugs
EHG2DVR3=Education FDCDVAVD=Food HWTGDBCC=BMI INCDGPER=Income OHT_005=OralH
SMKDGSTP=StopSmoking; run;

*** Loop to group "Valid skip", "don't know", "refused", "did not state"
etc., responses together. Also group oral health "(oralh)" into good/poor
categories. ;
data smoking.dataoralh; set smoking.data1;
array x education bmi diabetes oralh toothbrushing
dentalvisits food stopsmoking drugs;
do i=1 to dim(x);
if x[i]>=6 and x[i]<=9 then x[i]=.; end; drop i;
if alcohol>=96 then alcohol=.;
if income>=96 then income=.;
if oralh<=3 and oralh>=1 then oralh = 1;
else if oralh>3 and oralh<6 then oralh = 0;
run;

**** Make and assign value labels. Then, apply the labels;
proc format;
value sex_1 1="Male" 2="Female";
value age_1 1="12 to 14"
           2="15 to 17"
           3="18 and 19"
           4="20 to 24"
           5="25 to 29"
           6="30 to 34"
           7="35 to 39"
           8="40 to 44"
           9="45 to 49"
          10="50 to 54"
          11="55 to 59"
          12="60 to 64"
          13="65 to 69"
          14="70 to 74"
          15="75 to 79"
          16=">= 80";
value edu_1 1="< high school"
           2="Graduated high school, no post-secondary"
           3="Post-secondary diploma or degree";
value bmi_1 1="Underweight"
```

```

                2="Normal weight"
                3="Overweight"
                4="Obese - Class I, II, III";
value diab_1      1="Yes"
                  2="No";
value oralh_1 0="Poor"
                1="Good";
value teethfreq_1 1="At least 1x day"
                  2="At least 1x week"
                  3="At least 1x month"
                  4="At least 1x year";
value dentalvisit_1 1="> 1x a year"
                    2="About 1x a year"
                    3="< 1x a year"
                    4="Only for emergency care"
                    5="Never";
value food_1 1="Avoids foods - fat, salt, cholesterol, calories"
              2="Does not avoid foods - fat, salt, cholesterol, calories";
value smokestop_1 0="< 1 year"
                  1="1 to 2 years"
                  2="3 to 5 years"
                  3="6 to 10 years"
                  4=">= 11 years";
value alcohol_1 1="< 1x a month"
                 2="1x a month"
                 3="2 to 3x a month"
                 4="1x a week"
                 5="2 to 3x a week"
                 6="4 to 6x a week"
                 7="1x a day";
value drugs_1 1="Used drugs - 12 months"
               2="No drugs - 12 months";
value income_1 1="No income, or income loss"
                2="< $20,000"
                3="$20,000 to $39,999"
                4="$40,000 to $59,999"
                5="$60,000 to $79,999"
                6=">= $80,000" ; run;
data smoking.dataoralh; set smoking.dataoralh; format Alcohol alcohol_1.
Diabetes diab_1. DentalVisits dentalvisit_1. Toothbrushing teethfreq_1.
Age Age_1. Sex Sex_1. Drugs Drugs_1. Education edu_1. Food Food_1. BMI bmi_1.
Income income_1. OralH oralh_1. StopSmoking smokestop_1.; run;

*** Descriptive statistics of data (all categorical);
proc freq data=smoking.dataoralh; tables _ALL_/missing; run;

*** Dichotomize smoking status. Then quick bivariate analysis. . . ;
data modsmoke; set smoking.dataoralh;
if stopsmoking=. then stopsmoking2="Non-former"; else stopsmoking2 =
"former"; run;

```

```

*. . .1) between exposure status and all variables (minus the outcome);
proc freq data=modsmoke;
tables stopsmoking2*(age income alcohol diabetes dentalvisits toothbrushing
sex drugs education food bmi oralh)/nopercent nocol nofreq nocum missing
chisq; run;
*. . . 2) between outcome status and all variables (minus the exposure);
proc freq data=modsmoke;
tables oralh*(age income alcohol diabetes dentalvisits toothbrushing sex
drugs education food bmi stopsmoking)/nopercent nocol nofreq nocum missing
chisq; run;
*. . . 3) between the exposure and the outcome (using a crude log-binomial
model);
proc genmod data=smoking.dummies descending;
class oralh (param=ref ref="Poor") stopsmoking (ref="< 1 year");
model oralh=stopsmoking/ dist=bin link=log lrci; estimate "RR Good vs Poor"
stopsmoking 1-1
/exp;run;

```

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.1961	0.0516	-0.3192	-0.1135	14.46	0.0001
StopSmoking	1 to 2 years	1	-0.1224	0.1135	-0.3963	0.0699	1.16	0.2809
StopSmoking	3 to 5 years	1	-0.0663	0.1285	-0.4101	0.1286	0.27	0.6058
StopSmoking	6 to 10 years	1	-0.0046	0.1216	-0.3487	0.1755	0.00	0.9697
StopSmoking	>= 11 years	1	0.0544	0.0662	-0.0793	0.1943	0.68	0.4111
Scale		0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Contrast Estimate Results									
Label	Mean Estimate	Mean		L'Beta Estimate	Standard Error	Alpha	L'Beta		Chi-Square
		Confidence Limits					Confidence Limits		
RR Good vs Poor	0.9455	0.6385	1.4000	-0.0561	0.2003	0.05	-0.4487	0.3365	0.08
Exp(RR Good vs Poor)				0.9455	0.1894	0.05	0.6385	1.4000	

*** Non-significant.

*** Regardless, all other variables seem to be associated with the exposure and outcome (could be due to large sample size) from the Chi-squared test.

*** Trying log-binomial models of the other factors' associations with the outcome (first trying "age" modelled with outcome): ;

```

proc genmod data=smoking.dummies descending;
class oralh (param=ref ref="Poor")

```

```
age (ref="12 to 14");
model oralh=age/ dist=bin link=log lrci;
estimate "RR Good vs Poor" age 1-1/exp;run;
*** Proper entry of dummies for age? ;
```

Class Level Information														
Class	Value	Design Variables												
Age	12 to 14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	15 to 17	1	0	0	0	0	0	0	0	0	0	0	0	0
	18 and 19	0	1	0	0	0	0	0	0	0	0	0	0	0
	20 to 24	0	0	1	0	0	0	0	0	0	0	0	0	0
	25 to 29	0	0	0	1	0	0	0	0	0	0	0	0	0
	30 to 34	0	0	0	0	1	0	0	0	0	0	0	0	0
	35 to 39	0	0	0	0	0	1	0	0	0	0	0	0	0
	40 to 44	0	0	0	0	0	0	1	0	0	0	0	0	0
	45 to 49	0	0	0	0	0	0	0	1	0	0	0	0	0
	50 to 54	0	0	0	0	0	0	0	0	1	0	0	0	0
	55 to 59	0	0	0	0	0	0	0	0	0	1	0	0	0
	60 to 64	0	0	0	0	0	0	0	0	0	0	1	0	0
	65 to 69	0	0	0	0	0	0	0	0	0	0	0	1	0
	70 to 74	0	0	0	0	0	0	0	0	0	0	0	0	1
	75 to 79	0	0	0	0	0	0	0	0	0	0	0	0	1
	>= 80	0	0	0	0	0	0	0	0	0	0	0	0	1

```
*** Yes. But model non-convergent.
```

```
*** Try with Poisson instead-- Poisson can approximate binomial distribution;
data smoking.dummies; set smoking.dummies;
by age notsorted; if age then id+1;run;
proc genmod data=smoking.dummies; class id;
model oralh=age/ dist=poisson link=log lrci;
repeated subject=id;
estimate "RR Good vs Poor" age 1-1/exp;run;
*** We'll use p<.10 as our threshold for potential confounding. Age here not
"significantly" associated with the outcome (oralh), however.
```

```
*** Chisq again to see if oralh differs across only some ages: ;
proc freq data=smoking.dummies;
tables oralh*age/ chisq; run;
*** Some differences in oralh at the extremes of age. Testing if
dichotomizing age results in a "significant" model at p<.10: ;
data smoking.dummies2; set smoking.dummies;
if age<=8 then age_bin = 0;
else if age>8 then age_bin = 1; run;
proc genmod data=smoking.dummies2 descending;
class oralh (param=ref ref="Poor")
age_bin (ref="0");
model oralh=age_bin/ dist=bin link=log lrci;
estimate "RR Good vs Poor" age_bin 1-1/exp;run;
```

```
** Algorithm convergence. But p>.10. Leave age out of full model.
```

** Use genmod again to test all other remaining factors in log-binomial (or Poisson) models with oralh at p>.10 significance (potential confounders): ;

```
proc genmod . . .
```

** . . . only education, bmi, and drug use associated with oralh at p<.10. Place these factors in log-bin model with oralh and the exposure: ;

```
proc genmod data=smoking.dummies descending;
class oralh(param=ref ref="Poor")
drugs(ref="Used drugs - 12 months")
BMI(ref="Obese - Class I, II, III")
stopsmoking(ref="< 1 year");
model oralh=stopsmoking education bmi drugs / dist=bin link=log lrci;
estimate "RR Good vs Poor" stopsmoking 1-1/exp;run;
```

*** Non-convergence. Trying multivariate Poisson regression instead. ;

```
data smoking.dummies; set smoking.dummies;
by drugs notsorted; if drugs then id+1;run;
proc genmod data=smoking.dummies; class id;
model oralh=stopsmoking bmi drugs / dist=poisson link=log lrci;
repeated subject=id;
estimate "RR Good vs Poor" stopsmoking 1-1/exp;run;
```

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept	-0.5024	0.2816	-1.0543	0.0494	-1.78	0.0744
StopSmoking	-0.0218	0.0298	-0.0803	0.0366	-0.73	0.4643
BMI	-0.0188	0.0612	-0.1388	0.1012	-0.31	0.7592
Drugs	0.2601	0.1473	-0.0286	0.5487	1.77	0.0774

Contrast Estimate Results										
Label	Mean Estimate	Mean		L'Beta Estimate	Standard Error	Alpha	L'Beta		Chi-Square	Pr > ChiSq
		Confidence Limits					Confidence Limits			
RR Good vs Poor	0.9784	0.9229	1.0373	-0.0218	0.0298	0.05	-0.0803	0.0366	0.54	0.4643
Exp(RR Good vs Poor)				0.9784	0.0292	0.05	0.9229	1.0373		

*** Still not significant. But could try removing variables in order of their highest p-value (descending order) anyhow. Would check if the estimate (beta coefficient) for StopSmoking changes by ~±20%.

*** One could also compute changes in the mean squared error (MSE) of StopSmoking, as per the recent paper by Greenland et al. (2016).