Use Capture-Recapture Method To Estimate Prevalence Of Disease In SAS

Lin Yan[1,2]; Lisa M Lix[1,2]

1. Department of Community Health Sciences, University of Manitoba, Canada

2. George & Fay Yee Centre for Healthcare Innovation, University of Manitoba, Canada

INTRODUCTION

- Capture-recapture (CR) models, originally developed to estimate the size of animal populations, have been adapted for use by epidemiologists to estimate the total size of disease populations for such conditions as cancer, diabetes, and arthritis.
- Two assumptions, independence of capture in each data source and homogeneity of capture probabilities, which underlie conventional CR models, are unlikely to hold in epidemiological studies. Failure to satisfy these assumptions may bias population size estimates.
- Several statistical models have been proposed to incorporate dependency amongst sources and covariates to model heterogeneity in capture probabilities. However, none of these models is optimal and researchers may be unfamiliar with how to use them in practice.

OBJECTIVES

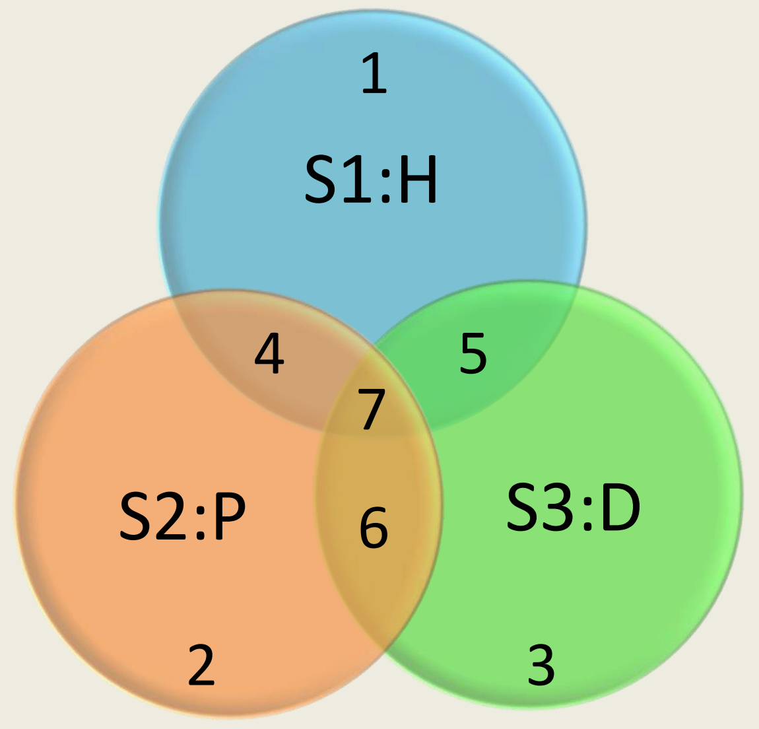
- To demonstrate the implementation of the log-linear (LL), multinomial logit (ML) and conditional logit (CL) CR models in SAS for estimating the number of missed disease cases from incomplete population-based data sources.
- To review advantages and disadvantages of each of these CR models.

A REAL WORLD EXAMPLE

- We demonstrate CR models for estimating the prevalence of Parkinson's Disease (PD) with administrative health data from the province of Saskatchewan.
- **Data Sources:** Hospital discharge abstracts (H), physician billing claims (P), and prescription drug records (D). All databases can be linked via a unique, anonymized personal health identifier that is found in the population registry.
- **Model:** We adopted a three-source CR model to illustrate each method.

METHODS

Table 1. Comparison of Three Capture-Recapture (CR) Models

	Log-Linear (LL) Model	Multinomial Logit (ML) Model	Conditional Logit (CL) Model
Model Definition	Define a contingency table in which all individuals are classified into *K* mutually-exclusive capture profiles (Figure 1) **Figure 1: Capture Profiles for Three-Source CR Model**  **1=Hospital Only** **2=Physician Only** **3=Drug Only** **4=Hospital & Physician** **5=Hospital & Drug** **6=Physician & Drug** **7=All Sources**	For the *i*th individual (*i*=1, …, *n*) , the probability of belonging to the *k*th capture profiles (*k*=1, …, *K*) is estimated by $\varPi_{ik} = \frac{\exp\{\eta_{ik}\}}{\sum_{r=1}^{K} \exp\{\eta_{ir}\}}$, where $\eta_{ik}$is the log-odds of the *k*th profile ($\eta_{ik=}Log\,\frac{\pi_{ik}}{\pi_{iK}}$) from a linear model. The probability of not being captured by any source can be estimated as $\varPi_{0\|i}= \frac{m_{0\|i}}{1+m_{0\|i}}$, where, $m_{0\|i}$ is the individual's contribution to the estimate of the missed number of cases	For the *i*th individual (*i*=1, …, *n*) define the ith row of the covariate vector **X** as **x**$_i$ with *H* elements. Each individual is classified in one of *K* unique capture profiles. The probability of not being captured by any source is: $\varPi_{0\|i} = \frac{1}{1+\sum_{r=1}^{K} \exp\left(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jr}\right)}$, where x_{ih} is the *h*th element (*h* = 1 ,…, *H*) of **x**$_i$, λ_{hj} is the *h*th element of the regression parameter matrix **Λ** (*H* × *J*), and y_{jr} is the *j*th element of the design matrix **Y** (*J* × *K*)
Dependence	Use interaction terms	$m_{0\|i}$ can be estimated based on the assumptions of sources dependency (see next section)	Formulate a design matrix which contains main and interaction effects
Heterogeneity	Stratify the contingency table	Include covariates in the model	Include covariates in the model
Estimate	Number of missed disease cases	Total population size: $N = \sum_{i=1}^{n}(\frac{1}{1-\varPi_{0\|i}})$, which includes observed and missed cases	Total population size: $N = \sum_{i=1}^{n}(\frac{1}{1-\varPi_{0\|i}})$, which includes observed and missed cases
Strengths and Limitations	Classical CR model. Easy to use. However, stratification by 2+ covariates can lead to small sample sizes and thus increase the variation of the population size estimates. Continuous covariates cannot be included	Both continuous and categorical covariates can be included. Dependence between sources can be considered. The data structure is simple. However, this method is only applicable to the three-source CR model	This model combines the LL and ML models, enables modeling of dependence between sources and different capture probabilities for each individuals. This method can be extended to more than three sources. However, coding and data manipulation are complicated.

LOG-LINEAR MODEL

Data Preparation

S1	S2	S3	Count
1	0	0	140
0	1	0	1601
0	0	1	2350
1	1	0	84
1	0	1	71
0	1	1	1604
1	1	1	297
0	0	0	.

Create a dataset for LL model analysis:

- S1: Captured by Hospital Data
- S2: Captured by Physician Claims Data
- S3: Captured by Drug Data
- For S1, S2 and S3, a code of 1 indicates captured and 0 not captured
- Last row (S1=0 and S2=0 and S3=0) represents the number of missed cases in all sources

Set this cell as a missing value, which will be estimated in the model step

SAS Code and Output

```
proc genmod data=pd_nocov_llm;
model count = S1 S2 S3 S1*S3 S1*S2/ dist = p obstats;
output out=pd_3s_nocov_L predicted=predicted;
run;
```

Model includes main effects and two-way interactions (S1 - S2; S1 - S3); the latter are used to account for dependencies bewteen sources

A separate CR model is estimated for each stratum to address heterogeneity of capture probabilities

Observation Statistics

Observation	count	S1	S2	S3	Predicted Value
1	140	1	0	0	79.837838
2	1601	0	1	0	1601
3	2350	0	0	1	2350
4	84	1	1	0	144.16216
5	71	1	0	1	131.16216
6	1604	0	1	1	1604
7	297	1	1	1	236.83784
8	.	0	0	0	2345.6047

Estimate the number of cases not captured in any data source

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	113.8227	113.8227
Scaled Deviance	1	113.8227	113.8227
Pearson Chi-Square	1	113.3206	113.3206
Scaled Pearson X2	1	113.3206	113.3206
Log Likelihood		38745.6195	
Full Log Likelihood		-84.2719	
AIC (smaller is better)		180.5438	
BIC (smaller is better)		180.2193	

The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) can be used to compare model fit

MULTINOMIAL LOGIT MODEL

Data Preparation

ID	Sex	Agecat	Source
1	1	1	1
2	2	2	2
3	1	1	3
…	…	…	…
N_i	2	1	3

Create a dataset for ML model analysis:

- Contains Individual-level data, one record per person.
- Source: a variable that has K mutually exclusive categories, where K is the number of capture profiles, in three sources analysis, K=7.
- Agecat: age group, 1 for '<65' and 2 for '65+'.
- Sex: 1 for male and 2 for female.

SAS Code And Output

```
proc logistic data = sgf.sgf_pd_3s;
class source (ref = "7") sex  / param = ref;
model source =sex agecat   / link = glogit scale=none;
output out=pd_cov_mlm_t p=PROB;
run;
```

Sex and age group are included in the model to address heterogeneity in capture probabilities

Output a dataset which contains probabilities of being captured for each individual

Capture Profile

Probability of Being Captured

ID	_LEVEL_	PROB
1	1	0.03764
1	2	0.27168
1	3	0.20465
1	4	0.02425
1	5	0.0181
1	6	0.36383
1	7	0.07986

```
proc transpose data= pd_cov_mlm_t out=pd_cov_mlm_t1
(drop=_label_ _name_) prefix=prob;
by studyid;
var p;
run;
```

Convert the dataset from a long to wide format

ID	prob1	prob2	prob3	prob4	prob5	prob6	prob7
1	0.03764	0.27168	0.20465	0.02425	0.0181	0.36383	0.07986

```
data pd_cov_mlm_t2;
set pd_cov_mlm_t1;
m0=(prob2*prob3)/(prob6);
p0=m0/(1+m0);
recip_P0=1/(1-P0);
run;
```

Calculate $m_{0|i}$; in this example, we assume that sources 1 and 2, and sources 1 and 3 are dependent. For calculating $m_{0|i}$ under other assumptions, see Table 2

Calculate the probability of not being captured by any source

Table 2. Method for Calculating $m_{0|i}$ Under Different Assumptions of Source Dependency

Interaction	$m_{0\|i}$
S1*S2	$(\pi_{i1} + \pi_{i2} + \pi_{i4})* \pi_{i3} /(\pi_{i7} + \pi_{i6} + \pi_{i5})$
S1*S3	$\pi_{i1} + \pi_{i3} + \pi_{i5})* \pi_{i2} /(\pi_{i7} + \pi_{i6} + \pi_{i4})$
S2*S3	$(\pi_{i6} + \pi_{i3} + \pi_{i2})* \pi_{i1} /(\pi_{i7} + \pi_{i5} + \pi_{i4})$
S1*S2; S1*S3	$(\pi_{i2} * \pi_{i3})/(\pi_{i6})$
S1 *S;, S2*S3	$(\pi_{i1} * \pi_{i3})/(\pi_{i5})$
S1*S3; S2*S3	$(\pi_{i1}* \pi_{i2})/(\pi_{i4})$

```
proc univariate data=pd_cov_mlm_t2;
var recip_p0;
Run;
```

The UNIVARIATE Procedure (Variable: recip_P0)

Moments

N	6147	Sum Weights	6147
Mean	1.4675198	Sum Observations	9020.84422
Std Deviation	0.33266871	Variance	0.11066847

Estimated number of cases = Observed + Missed

CONDITIONAL LOGIT MODEL

Data Preparation

Design Matrix

	Capture Profile						
	100	010	001	110	101	011	111
L1	1	0	0	1	1	0	1
L2	0	1	0	1	0	1	1
L3	0	0	1	0	1	1	1
L1_L2	0	0	0	1	0	0	1
L1_L3	0	0	0	0	1	0	1

Create a Design Matrix based on a model assuming that source 1 and 2 and source 1 and 3 are dependent.

- **The labels for the rows: main effects (L1, L2, and L3) and interaction effects (L1_L2 and L1_L3).**
- **The labels for the columns: capture profiles.**
- **The design matrix was coded with "L" for discriminating with variables (S1, S2, and S3) that already existed in the original dataset.**

Structure of Dataset

ID	CP	S_YN	L1	L2	L3	L1_L2	L1_L3	Sex	agecat
1	1	0	1	0	0	0	0	1	1
1	2	0	0	1	0	0	0	1	1
1	3	0	0	0	1	0	0	1	1
1	4	0	1	1	0	1	0	1	1
1	5	0	1	0	1	0	1	1	1
1	6	1	0	1	1	0	0	1	1
1	7	0	1	1	1	1	1	1	1

Create a dataset for CL model :

- **Contains individual-level data; 7 records per person**
- **Agecat: age group; 1 for '<65' and 2 for '65+'.**
- **CP: Capture Profiles**
- **Dependent variable(S_YN): for each capture patterns, 1 = yes; 0 = no**

SAS Code And Output

```
proc logistic data=pd_data_clm outest=parameters_cov_c;
model S_YN (ref='0')= L1 L2 L3 L1_L2 L1_L3
L1*sex L2*sex L3*sex L1_L2*sex L1_L3*sex
L1*agecat L2*agecat L3*agecat L1_L2*agecat L1_L3*agecat/
link=logit ;
strata studyid;
run;
```

Model include main effects and interaction effects based on assumption of source dependency, as well as covariates for dealing with heterogeneity

Table of parameter estimates

Analysis of Conditional Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
L1	1	-8.6467	0.5423	254.2520	<.0001
L2	1	-2.8091	0.1373	418.7664	<.0001
L3	1	-1.1349	0.1499	57.3440	<.0001
L1L2	1	3.9504	0.5211	57.4789	<.0001
L1L3	1	1.6622	0.5002	11.0420	0.0009
L1*sex	1	0.8776	0.1958	20.0888	<.0001
L2*sex	1	0.8167	0.0680	144.3406	<.0001
L3*sex	1	0.1494	0.0716	4.3492	0.0370
L1L2*sex	1	-0.5995	0.1859	10.3944	0.0013
L1L3*sex	1	-0.2331	0.1856	1.5774	0.2091
L1*agecat	1	1.9376	0.1903	103.6723	<.0001
L2*agecat	1	0.8560	0.0514	277.7278	<.0001
L3*agecat	1	0.4259	0.0552	59.5559	<.0001
L1L2*agecat	1	-1.1013	0.1826	36.3658	<.0001
L1L3*agecat	1	-0.4201	0.1756	5.7250	0.0167

```
proc iml;
varNames={L1 L2 L3 L1_L2 L1_L3 L1sex L2sex L3sex
L1_L2sex  L1_L3sex L1agecat L2agecat L3agecat
L1_L2agecat L1_L3agecat};
use parameters_cov_c; read all var varNames into lambda;
close parameters_cov_c;
matrix_lambda = shape(lambda, 3, 5);

varNames={Intercept sex agecat};
use data_cov; read all var varNames into X; close data_cov;
use Design_M; read all var _all_ into Y; close Design_M;

A=X*Matrix_lambda*Y; B=exp(A); P0=1/(1+B[,+]);

recip_P0=1/(1-P0); nNames="recip_P0";
create cov_clm from recip_P0[colname=nNames];
append from recip_P0;
run;

proc univariate data=cov_clm; var recip_p0; run
```

Create Λ from estimates of model parameters

-8.64671	-2.8091	-1.13489	3.950435	1.662243
0.877563	0.816662	0.149381	-0.59945	-0.2331
1.937641	0.855977	0.425868	-1.10131	-0.42009

Matrix A $\begin{bmatrix} 1 & x_{sex1} & x_{age1} \\ 2 & x_{sex2} & x_{age2} \\ \vdots & \vdots & \vdots \\ n & x_{sexn} & x_{agen} \end{bmatrix} * \begin{bmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} & \lambda_{14} & \lambda_{15} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} & \lambda_{24} & \lambda_{25} \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & \lambda_{34} & \lambda_{35} \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$

X Λ Y

The UNIVARIATE Procedure (Variable: recip_P0)			
Moments			
N	6147	Sum Weights	6147
Mean	1.467527	Sum Observations	9020.891
Std Deviation	0.332685	Variance	0.110679

Estimated number of cases = Observed + Missed

SUMMARY

Table 3. Comparison of Results from Three CR Models

Interaction	Estimated Number of Missed Cases					
	Intercept Only Model			Covariates Model*		
	LL	ML	CL	LL	ML	CL
S2*S3	1720	1720	1720	1718	1692	1679
S1*S3	2065	2065	2065	3301	2586	2560
S1*S2	2174	2174	2174	3433	2681	2669
S1*S3, S2*S3	2668	2668	2668	2608	2617	2617
S1*S2, S2*S3	4633	4633	4633	4677	6007	6009
S1*S2, S1*S3	2345	2345	2345	3744	2873	2874

*Covariates: age group; sex

- For models without covariates, the estimated number of missed PD cases are the same for LL, ML and CL model
- For models including the covariates of age group and sex, the ML and CL models produced similar estimates of the number of missed PD cases, while the LL model predicted a much higher number of missed cases. This latter result may be due to small sample size in some cells of contingency table due to stratification

REFERENCES

Zwane, E. & van der Heijden, P. (2005). Population estimation using the multiple system estimator in the presence of continuous covariates. *Statistical Modelling, 5,* 39-52.

Rossi, G., Pepe, P., Curzio, O., & Marchi, M. (2010). Generalized linear models and Capture-Recapture Method in a closed population: strengths and weaknesses. *Statistica, 70*(3), 371-390.

ACKNOWLEDGEMENTS

This study was supported by funding from the Canadian Institutes of Health Research (Funding Reference No. 123357) and the Saskatchewan Health Research Foundation. This research is based in part on de-identified data provided by the Saskatchewan Ministry of Health. The interpretation and conclusions contained herein do not necessarily represent those of the Government of Saskatchewan or the Ministry of Health.

For further information : Dr. Lin Yan; lin.yan@umanitoba.ca