

Coding IPW and SMR in SAS and Stata

Bailey M. DeBarmore

Suggested Citation:

DeBarmore BM. "Coding IPW and SMR in SAS and Stata". 2019. [PDF File]. Retrieved from <http://www.baileydebarmore.com/epicode/calculating-ipw-and-smr-in-sas>



Table of contents

Preface	3
Weighting vs regression	4
Propensity scores	5
Standardized Mortality Ratios (SMR)	8
Calculating SMR in SAS	11
Calculating SMR in Stata	12
Inverse Probability of Treatment Weights (IPTW)	13
Calculating IPTW in SAS	15
Calculating IPTW in Stata	18

Preface

This guide is meant to walk you through the basic “why” we might use propensity scores (inverse probability weights and standardized mortality/morbidity ratios) and then jump into the “how”.

Learning about a method in class is different than implementing in practice.

If you’re like me, the mathematical notation doesn’t usually make the leap from class to manuscript, so I’m going to highlight the methodologic considerations you do need to remember when choosing between IPW and SMR and other methods, and then show you how to code them in SAS and Stata.

There are many ways to teach statistics and epidemiology, and the examples I put forth is just my way. If you find yourself wanting to add additional details to my explanations, fantastic! You have a mind for the mathematical side. Help your peers by teaching them, because when you teach you learn more.

If these methods are a struggle for you, don’t fret. Though statisticians loathe the “What test to use?” decision tree, if you are interested in very applied methods, don’t shy away from adding these methods to your toolbox and moving on. I suggest as you go through this guide, writing out a summary in your own words that you can reference later. Even better, copy and paste that summary into your code, and save it very inconspicuously.

This guide is intended for everyone – from high school students and undergraduates interested in the world of statistics and epidemiology, to professors teaching the next generation of epidemiologists.

Let’s get started!

Bailey DeBarmore

Weighting versus Regression

Effect estimate interpretations when you use weighting are marginal effects in the target population.

When you adjust for covariates in a regression model, you are interpreting a conditional effect, that is, the effect of the exposure holding the other covariates being constant (adjusting for).

Why do we care about marginal versus conditional?

Conditional estimates are troublesome with time-varying covariates, because we run into collider bias and conditioning on mediators (a big no-no), thus in those situations, weights are preferable to throwing all of our confounders in a model.

Even in simpler situations, without time-varying covariates, using weights over multivariable regression can help with convergence issues, like if you want to estimate risk differences but your model won't converge.

Propensity Scores

A propensity score is a predicted probability that may be used to predict exposure (or treatment) status, but can also be used for censoring or missingness.

How do we use propensity scores for confounding?

We can use propensity scores to generate weights, which when applied to the final model, make the exposure independent from confounders (Figure 1).

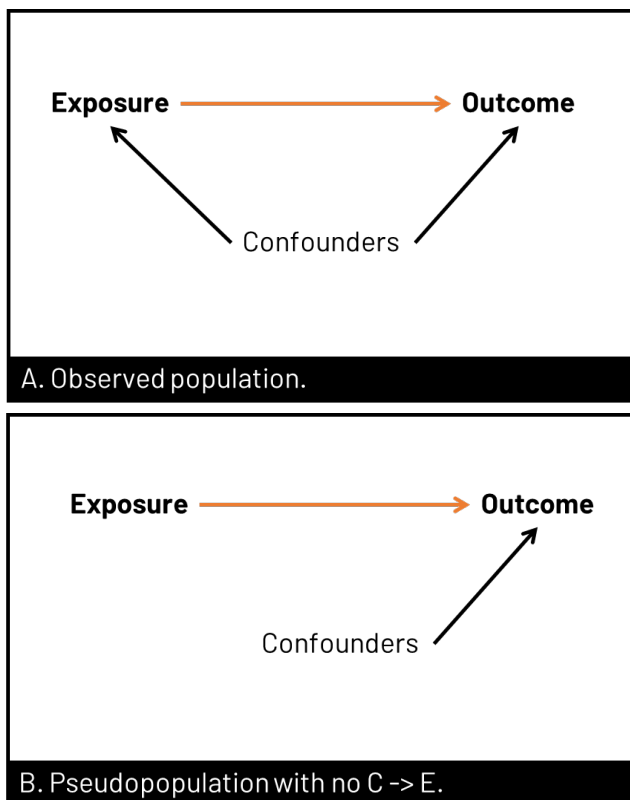



Figure 1. Panel A shows the observed data, where the relationship between exposure and outcome is confounded by, well, confounders. In panel B, we have removed the arrow from confounders to exposure. We can remove the arrow in several ways, including using propensity scores (of various types) to create a pseudopopulation where exposure and confounder are no longer associated.

Abstractly speaking, we remove the arrow from confounders to exposure by crafting a pseudopopulation. Logistically speaking, we model the association between exposure and confounders (Figure 2). Contrast that with the main analysis where we model the outcome and exposure.

$$\text{Outcome} = \text{Exposure} + X1 + X2 + X3$$


confounders

A. Typical multivariable model for relationship of interest

$$\text{Exposure} = X1 + X2 + X3$$


confounders

B. Multivariable model to create pseudopopulation

Figure 2. Panel A shows the usual multivariable model we run in our analyses – to estimate the association of the exposure with the outcome, controlled for confounders. When we want to use propensity scores, first we create the weights that we will later use in our final model, by modeling the association of the confounders with the exposure – so we can remove that arrow like in Figure 1B.

Propensity scores can also control for confounding via covariate adjustment (I discourage you from this option), stratification, and matching, in addition to weighting.

Standardized Mortality Ratios (SMR)

SMR stands for standardized mortality/morbidity ratios and are, at their core, weights.

However, you'll often see SMR in the context of indirect standardization. Standardization is a form of weighting data to look more similar to another set of data, so that you can compare them without the event rates being confounded like age.

In this guide, I'll be talking about SMR as weights to apply in your regression.

Inverse Probability Weights (IPW)

IPW stands for inverse probability weights. They are a general type of propensity score that can be used to address confounding as well as censoring.

When used for confounding, they are called inverse probability of treatment weights (IPTW) even if you're talking about an exposure and not an assigned treatment.

We'll contrast the pseudopopulation you create with IPW versus SMR, as well as touch on unstabilized versus stabilized IPTWs.

SMR

The key thing to remember with SMR weights is that you're estimating the **average treatment effect in the treated** (also called the **average exposure effect in the exposed**).

In other words, you estimate the effect had the exposed group been exposed (observed) versus had the exposed group been unexposed. After you read through the IPW section, come back here and re-read so that you can understand SMR in contrast with IPW.

The pseudopopulation that you create has a covariate (confounder) distribution equal to that observed in the exposed group (Figure 3A).

You can also generate your SMR weights where the unexposed group is the target of interest, and model the effect had the unexposed group been unexposed versus the unexposed group been exposed. What covariate distribution will you use? The covariate distribution of the unexposed group (Figure 3B).

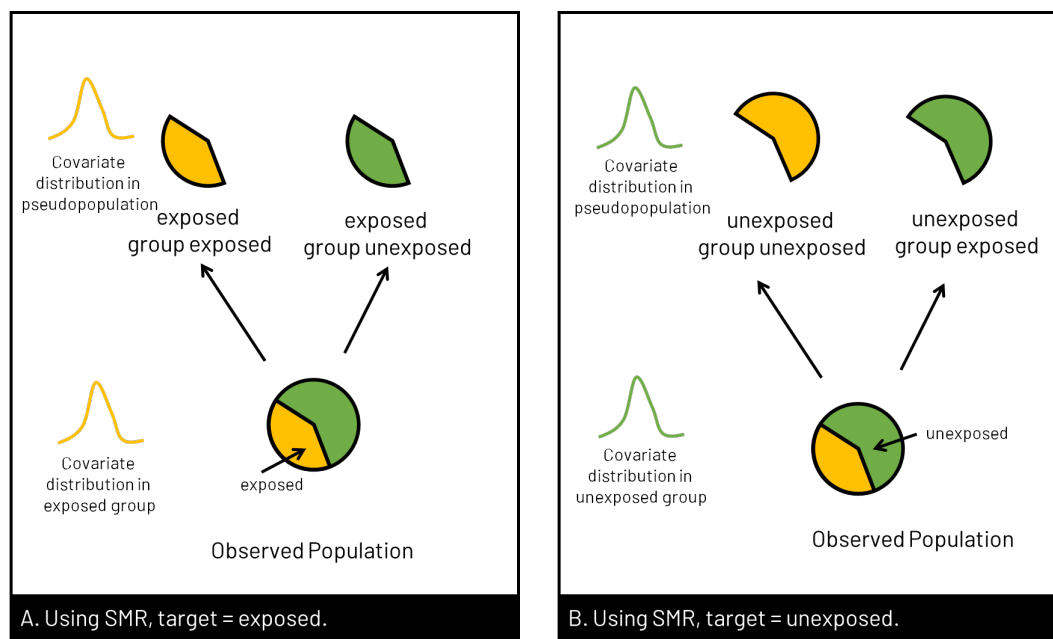


Figure 3. In Panel A, the target group is the exposed group, so we use SMR to model the counterfactual – had the exposed group been unexposed – versus what we observed (the exposed group exposed).

Calculating SMR

While the sight of probabilities might give you unwanted flashbacks to introductory statistics, don't skip over this section.

Understanding the weights we calculate for each of the scenarios on the previous page are instrumental in understanding how we calculate the weights in SAS. In Stata, the program does it behind the scenes for you.

If we think about exposure or treatment assignment as A, then in the exposed group A=1, and in the unexposed group, A=0. If we think of the covariate distribution as Z, we will always note Z=z, that is, the covariate distribution equals what we observe in that group.

The SMR weight for the target group will always equal 1. The weight for the other group will be the probability of the target group over the probability of the other group. It makes more sense below.

Weight (exposed)= <i>A=1, covariate dist Z=z</i>	$\frac{P(A = 1 Z = z)}{P(A = 1 Z = z)}$
Weight (unexposed)= <i>A=0</i>	$\frac{P(A = 1 Z = z)}{P(A = 0 Z = z)}$

A. SMR weights, target = exposed.

Can you write out the probabilities to calculate weights when the target group is the unexposed? (Answers next page)

Figure 4A. When the target group is the exposed group, the weight will equal zero because we are dividing the probability of A=1 over the probability of A=1. For the unexposed group (the other group), we want the SMR weight to apply the covariate distribution of the exposed group, so we model the probability of A=1 over the probability of A=0 (what we observe).

$$\text{Weight (unexposed)} = \frac{P(A = 0 | Z = z)}{P(A = 0 | Z = z)}$$

A=0 covariate dist Z=z

$$\text{Weight (exposed)} = \frac{P(A = 0 | Z = z)}{P(A = 1 | Z = z)}$$

A=1

B. SMR weights, target = unexposed.

Figure 4B. When the target group is the unexposed group, the weight will equal zero because we are dividing the probability of A=0 over the probability of A=0. For the exposed group (the other group), we want the SMR weight to apply the covariate distribution of the unexposed group, so we model the probability of A=0 over the probability of A=1 (what we observe).

Coding

Are you ready to code the SMR weights?

On the next few pages I have the code for SAS and Stata. You can download the program code as a .txt file at www.github.com/baileydebarmore/epicode.git.

What is &let?

So that you can easily adapt my SAS code, I use &let statements at the beginning of my code blocks. After the equals sign, you would replace <data> with your dataset name, <exposure> with your exposure variable, and <outcome> with your outcome variable. The code as written will then run with those chosen variables. Note that you do need to replace <covariates> in the model statement with your confounders.

If you don't want to use &let statements, simply go through the code and anywhere you see &<text>, replace both the & and the text with your regular code.

Calculating SMR in SAS

```
*****
* Calculating SMR weights
*****;

&let data=<data>;
&let y=<outcome>;
&let x=<exposure>;
&let id=<id>;

*Estimate the predicted probability given
covariates;

proc logistic data=&data desc;
    model &x=<covariates>;
    output out=pred p=p1;
run;


*Generate the weights by exposure status,
for exposed group = target their weight
will be 1;

data <newdata>;
    set pred;
    p0 = 1-p1;
    odds = p1/p0;

    if &x=1 then wt=1;
    else wt=odds;
run;

*Final weighted analysis;

proc logistic data=<newdata> desc;
    weight wt;
    model &y = &x;
run;
```

$$\text{Exposure} = X1 + X2 + X3$$


confounders

Calculating SMR in Stata

You'll be using the built-in `teffects` command with options to specify SMR versus IPW in Stata. Since we want the average treatment effect on the treated (aka the SMR), we'll use the option **atet**.

```
*****
```


```
* Calculating SMR weights
```

```
*****;
```

```
* Syntax for teffects statement
```

```
*teffects ipw (<outcome>) (<exposure>  
    <covariates>), atet
```

Write the command
all on one line in
your code



```
*where <outcome> is your outcome variable, <exposure>  
is your exposure variable, and <covariates> is a list  
of your covariates to generate your weights.
```

```
*Example: Binary
```

```
*Outcome = lowbirthwt
```

```
*Exposure = maternalsmoke
```

```
*Covariates = maternalage nonwhite
```

```
*Use the teffects statement to generate your weights  
and then apply them in a logistic (default) model all  
in 1 step
```

```
teffects ipw (lowbirthwt) (maternalsmoke maternalage  
    nonwhite), atet
```

```
*If your outcome is continuous, you can specify a  
probit model
```

```
*Example: Continuous
```

```
*Outcome = birthwt
```

```
*Exposure = maternalsmoke
```

```
*Covariates = maternalage nonwhite
```

```
teffects ipw (birthwt) (maternalsmoke maternalage  
    nonwhite, probit), atet
```

IPTW

In contrast to SMR weights, when you use IPTW weights you are estimating the **average treatment effect**, which is the treatment effect in a study population with a covariate distribution equal to the entire observed study population (not just the exposed or unexposed).

In other words, you're modeling the complete counterfactual. You're estimating the effect of the exposure or treatment had the entire population been exposed versus had the entire population been unexposed.

Unstabilized IPTW are calculated by taking the inverse of the probability of exposure given the observed covariates (Figure 5A).

Stabilized IPTW have a numerator equal to the probability of observing that exposure, which "stabilizes" the weight.

How do these equations compare to the SMR equations?

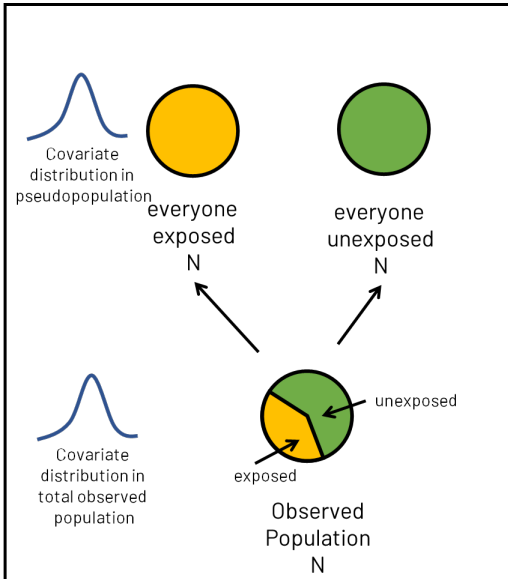
Weight (exposed)= $A=1$	$\frac{1}{P(A=1 Z=z)}$
Weight (unexposed)= $A=0$	$\frac{1}{P(A=0 Z=z)}$
<i>Note that $P(A=0 Z=z)=1-[P(A=1 Z=z)]$</i>	
A. Unstabilized IPTW.	

Weight (exposed)= $A=1$	$\frac{P(A=1)}{P(A=1 Z=z)}$
Weight (unexposed)= $A=0$	$\frac{P(A=0)}{P(A=0 Z=z)}$
<i>Note that $P(A=0)=1-[P(A=1)]$ and that $P(A=0 Z=z)=1-[P(A=1 Z=z)]$</i>	
B. Stabilized IPTW.	

Figure 5. Probability equation for unstabilized IPTW and stabilized IPTW.

On the next page we'll compare unstabilized and stabilized weights in terms of the pseudopopulation they create and the covariate distribution they apply.

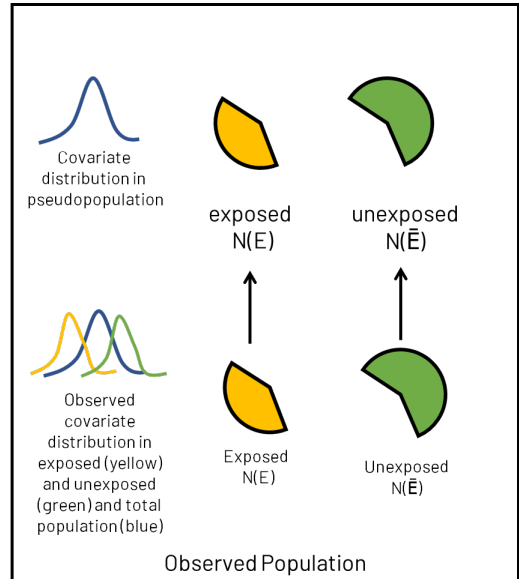
Unstabilized IPTW



A. Using unstabilized IPTW.

When you use unstabilized weights, you estimate the covariate distribution in the entire observed population (regardless of exposure status), and then create weights that apply that distribution to a pseudopopulation *twice* the size of your observed population. See in Panel A above how the exposed and unexposed groups have gone from partial circles to full circles?

Stabilized IPTW



B. Using stabilized IPTW.

When you use stabilized weights, you adjust the covariate distribution within exposure strata to match the overall covariate distribution. Because you are upweighting and downweighting people in their respective exposure group, we keep the population size the same.

Looking now at the covariate distributions, how do the images for unstabilized IPTW and stabilized IPTW compare to SMR?

With SMR, we match the covariate distribution of our “other group” to the distribution of our covariate group. This gives us the average treatment effect in the treated. In contrast with IPTW, we are using the covariate distribution of the overall population, which gives us the average treatment effect.

Calculating IPTW in SAS

```
*****
* Calculating IPTW
*****;
```

```
&let data=<data>;
&let y=<outcome>;
&let x=<exposure>;
&let id=<id>;
```

*Estimate denominator - output a dataset with results of regression called denom, with the resulting probabilities stored in variable d;

```
proc logistic data=&data desc;
    model &x = <covariates>;
    output out=denom p=d;
run;
```

$$P(A = 1 | Z = z)$$

$$P(A = 0 | Z = z)$$

*Generate numerator for stabilized weights - output a dataset with results of regression called num, with the resulting probabilities stored in variable n - note that there is nothing on the right side of the equation because the numerator will simply be $P(A=a)$, where a = observed exposure status;

```
proc logistic data=&data desc;
    model &x=;
    output out=num p=n;
run;
```

$$P(A = 1)$$

$$P(A = 0)$$

Weight (exposed)= A=1	$\frac{1}{P(A=1 Z=z)}$
Weight (unexposed)= A=0	$\frac{1}{P(A=0 Z=z)}$
<small>Note that $P(A=0 Z=z)=1-[P(A=1 Z=z)]$</small>	
A. Unstabilized IPTW.	

Weight (exposed)= A=1	$\frac{P(A=1)}{P(A=1 Z=z)}$
Weight (unexposed)= A=0	$\frac{P(A=0)}{P(A=0 Z=z)}$
<small>Note that $P(A=0)=1-[P(A=1)]$ and that $P(A=0 Z=z)=1-[P(A=1 Z=z)]$</small>	
B. Stabilized IPTW.	

*Generate stabilized and unstabilized weights by merging the datasets with regression output (merge on the unique identifier in your dataset, &id);

```
data <newdata>;
    merge &data denom num;
    by &id;
if &x=1 then do;
    uw = 1/d;
    sw = n/d;
end;
```

Creates stabilized and unstabilized weights for exposed group

*Remember we can use $1 - P(\text{exposed})$ for the unexposed weight components;

```
else if &x=0 then do;
    uw=1/(1-d);
    sw=(1-n)/(1-d);
end;
run;
```

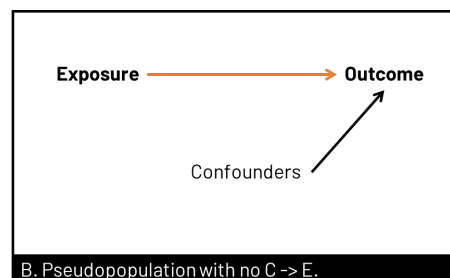
Creates stabilized and unstabilized weights for unexposed group

*Check the distribution of your IPTW - the mean should be 1. Is the sum for uw twice the sum of sw? why? is the range of uw greater than sw? why?;

```
proc means data=<newdata> mean sum min max;
var uw sw;
run;
```

*You can check to see if your exposure and covariates are associated in your new pseudopopulation (<newdata>);

```
proc logistic data=<newdata> desc;
    weight sw;
    model &x=<covariates>;
run;
```



B. Pseudopopulation with no C -> E.

*Now you can run your main analyses and apply the weights using the weight statement - use sw variable for stabilized weights, and use uw for unstabilized weights - you can use proc genmod, glm, logistic, etc. I'll show you below with logistic you can see how we're using &y and &x - and we don't need the covariates because the confounder -> x arrow is encompassed in the sw weight statement;


```
proc logistic data=<newdata> desc;  
    weight sw;  
    model &y = &x;  
run;
```

Calculating IPW in Stata

You'll be using the built-in `–teffects–` command with options to specify IPW versus SMR in Stata. Since we want the average treatment effect overall, we'll use option **ate** .

```
*****  
* Calculating SMR weights  
*****;  
* Syntax for teffects statement  
  
*teffects ipw (<outcome>) (<exposure>  
    <covariates>), ate
```

Write the command
all on one line in
your code



*where <outcome> is your outcome variable, <exposure> is your exposure variable, and <covariates> is a list of your covariates to generate your weights.

```
*Example: Binary  
*Outcome = lowbirthwt  
*Exposure = maternalsmoke  
*Covariates = maternalage nonwhite
```

*Use the `teffects` statement to generate your weights and then apply them in a logistic (default) model all in 1 step

```
teffects ipw (lowbirthwt) (maternalsmoke maternalage  
    nonwhite), ate
```

*If your outcome is continuous, you can specify a probit model

```
*Example: Continuous  
*Outcome = birthwt  
*Exposure = maternalsmoke  
*Covariates = maternalage nonwhite
```

```
teffects ipw (birthwt) (maternalsmoke maternalage  
    nonwhite, probit), atet
```