# Introducing a SAS® macro for doubly robust estimation

[1]Michele Jonsson Funk, PhD, [1]Daniel Westreich MSPH, [2]Marie Davidian PhD, [3]Chris Weisen PhD

[1]Department of Epidemiology and [3]Odum Institute for Research in Social Science, University of North Carolina, Chapel Hill, NC, USA.

[2]Department of Statistics, North Carolina State University, Raleigh, NC, USA.

## ABSTRACT

Estimation of the effect of a treatment or exposure with a causal interpretation from studies where exposure is not randomized may be biased if confounding and selection bias are not taken into appropriate account. Such adjustment for confounding is often carried out through regression modeling of the relationships among treatment, confounders, and outcome. Correct specification of the regression model is one of the most fundamental assumptions in statistical analysis. Even when all relevant confounders have been measured, an unbiased estimate of the treatment effect will be obtained only if the model itself reflects the true relationship among treatment, confounders, and the outcome. Outside of simulation studies, we can never know whether or not the model we have constructed includes all relevant confounders and accurately depicts those relationships. Doubly robust estimation of the effect of exposure on outcome combines inverse probability weighting by a propensity score with regression modeling in such a way that as long as *either* the propensity score model is correctly specified *or* the regression model is correctly specified the effect of the exposure on the outcome will be correctly estimated, assuming that there are no unmeasured confounders. While several authors have shown doubly-robust estimators to be powerful tools for modeling, they are not in common usage yet in part because they are difficult to implement. We have developed a simple SAS® macro for obtaining doubly robust estimates. We will present sample code and results from analyses of simulated data.

## INTRODUCTION

Correct specification of the regression model is one of the most fundamental assumptions in statistical analysis. Even when all relevant confounders have been measured, an unbiased estimate will be obtained only if the model itself reflects the true relationship among treatment, confounders, and the outcome. Outside of simulation studies, we can never know whether or not the model we have constructed accurately depicts those relationships. So the correct specification of the regression model is typically an unverifiable assumption.

Doubly robust (DR) estimation builds on the propensity score approach of Rosenbaum & Rubin (1983) and the inverse probability of weighting (IPW) approach of Robins and colleagues (Robins, 1998; Robins, 1998a; Robins, 1999; Robins, 1999a; Robins, Hernan, and Brumback, 2000). DR estimation combines inverse probability weighting by a propensity score with regression modeling of the relationship between covariates and outcome in such a way that as long as *either* the propensity score model *or* the regression model is correctly specified, the effect of the exposure on the outcome will be correctly estimated, assuming that there are no unmeasured confounders (Robins, Rotnitzky, and Zhao, 1994; Robins, 2000; van der Laan and Robins, 2003; Bang and Robins 2005). Specifically, one estimates the probability that a particular patient receives a given treatment as a function of that individual's covariates (the propensity score). Each individual observation is then given a weight equal to the inverse of this propensity score to create two pseudopopulations of exposed and unexposed subjects that now represent what would have happened to the entire population under those two treatment conditions. Maximum likelihood regression is conducted within these pseudopopulations with adjustment for confounders and risk factors. Results from extensive simulations by Lunceford and Davidian (2004) as well as Bang and Robins (2005) confirm the theoretical properties of this estimator.

## MATHEMATICS OF DOUBLY ROBUST ESTIMATION

We use the following notation: $Y$ is the observed response or outcome, $Z$ is a binary treatment (exposure) variable, and $\boldsymbol{X}$ represents a vector of baseline covariates. $Y_1$ and $Y_0$ are the counterfactual responses under treatment and no treatment, respectively. All of these variables are further subscripted by $i$ for subjects $i$=1, . . ., $n$. In this example, the causal effect of interest is the difference in means if everyone in the population received treatment versus everyone receiving no treatment, or $E(Y_1)-E(Y_0)$. In the following equation, $e(X,ß)$ is a postulated model for the true propensity score (from logistic regression) and $m_0(X,a_0)$ and $m_1(X,a_1)$ are postulated regression models for the true relationship between the vector of covariates (confounders plus other prognostic factors) and the outcome within each strata of treatment. With these definitions, the estimator of the causal effect is:

$$
\begin{aligned}
\widehat{\Delta}_{DR} &= n^{-1}\sum_{i=1}^{n}\left[\frac{Z_iY_i}{e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})} - \frac{\{Z_i - e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})\}}{e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})}m_1(\boldsymbol{X}_i,\widehat{\boldsymbol{\alpha}}_1)\right] \\
&\quad -n^{-1}\sum_{i=1}^{n}\left[\frac{(1-Z_i)Y_i}{1-e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})} + \frac{\{Z_i - e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})\}}{1-e(\boldsymbol{X}_i,\widehat{\boldsymbol{\beta}})}m_0(\boldsymbol{X}_i,\widehat{\boldsymbol{\alpha}}_0)\right] \\
&= \widehat{\mu}_{1,DR} - \widehat{\mu}_{0,DR}
\end{aligned}
$$

The standard error of is estimated using the empirical sandwich method (Lunceford & Davidian, 2004, equation 21; Stefanski & Boos 2002). Specifically, the sampling variance for the doubly robust estimator is calculated as $n^{-2}\sum_{i=1}^{n}\hat{I}_i^2$ where:

$$
\hat{I}_{DR,i} = \frac{Z_iY_i - m_1(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_1)(Z_i-\hat{e}_i)}{\hat{e}_i} - \frac{(1-Z_i)Y_i + m_0(\mathbf{X}_i,\hat{\boldsymbol{\alpha}}_0)(Z_i-\hat{e}_i)}{(1-\hat{e}_i)} - \hat{\Delta}_{DR}
$$

**IMPLEMENTING THE DR MACRO**

The DR macro runs two sets of models: one for the probability of receiving a dichotomous treatment or exposure and another to predict either the probability of the outcome (for a dichotomous outcome) or its mean value (for a continuous outcome) within strata of the exposure. We will introduce the macro using a simple example where the exposure of interest is statin use (statin) and the outcome of interest is risk of acute myocardial infarction (acuteMI) and two potential confounders (sex, age) have been measured:

```
%dr(%str(options data=cvdcohort descending;
  wtmodel statin = sex age / method=dr dist=bin showcurves;
  model acuteMI = sex age / dist=bin ;));
```

After the usual SAS® options to indicate the location of the dataset and the 'descending' option (which both function as they do in the GENMOD procedure), there are two model statements. The weight model (`wtmodel`) is the model for the propensity score or probability of treatment given covariates. The second model statement (`model`) is used to specify the covariates to be used for adjustment of the outcome regression model; in this case, risk of acute myocardial infarction adjusted for sex and age. Note that the main exposure (`statin`) is NOT included in the second model statement because these regression models are performed within the treatment groups. For the outcome regression models, the treatment groups have been reweighted such that they represent the expected outcomes under the two treatment conditions: if all individuals had been treated and then as if all individuals had been untreated. To describe the macro syntax, we show a more general implementation of the doubly robust macro with a binary exposure (binexp), three covariates for adjustment (x1 x2 x3) and a continuous outcome (response1).

```
%dr(%str(options data=sim descending;
 wtmodel binexp = x1 x2 x3 / method=dr dist=bin showcurves;
 model response1= x1 x2 x3 / dist=n;));
```

`%dr` is the name of the macro.

`(%str(options` is required code.

`data=`*`SAS-data-set`* is the standard SAS® method for indicating which dataset should be used for the analysis.

`descending` is standard SAS® coding so that the macro models the probability of the higher response value rather than lower response value (the default behavior); if your exposure and outcome are coded as 0=no / 1=yes, you want to be sure to include this option. Currently, invoking the `descending` option applies this to both the propensity score model (`wtmodel`) and the outcome regression model (`model`) when both are dichotomous variables.

**PROPENSITY SCORE MODEL**
The propensity score model is specified in the first model statement using the form:

`wtmodel` *`exposure = <covariates> </options>`*

We model the main exposure or treatment on the left side of the equals sign as a function of the covariates on the right side. It is appropriate to include as covariates all confounders as well as risk factors for the outcome of interest. Options should be specified after a slash (/) :

`method=dr` indicates that the doubly robust estimation method should be used. In the future, there will be other methods that the user can specify to obtain a propensity score adjusted estimate or a standard (not doubly-robust) inverse probability of treatment weighted estimate.

`dist=bin` indicates that this is a logistic regression (log-binomial will be coming in the future, at which point a "link="option will also be available). This is currently required; will be changed in the future to allow for option of "n" for a normal distribution, if the main exposure is continuous.

`showcurves` will produce the graphs of the two overlapping propensity score curves.

`common_support=`*`number`* where number is between 0 and 1 (inclusive), will limit the region of analysis to those observations for which there is common support for counterfactual inference. A value of 1 indicates that the program should use the entire region of common support; a value of 0.8 would indicate that the program should use only the middle 80% region of common support. In addition, the region of common support will be shown as vertical lines in the plot generated by `showcurves`.

**OUTCOME REGRESSION MODEL**
The outcome regression model is specified in the second model statement using the form:

`model` *`outcome =<covariates> </options>`*

This models the main outcome of interest (continuous or dichotomous) as a function of the covariates. The main exposure should NOT be specified here a second time. Options should be specified after a slash (/):

`dist=n` indicates that the model is a linear regression, with a normal error distribution. This is appropriate for a continuous outcome variable. If the outcome variable is dichotomous, `dist=bin` is appropriate.

## OUTPUT
Currently, the DR macro outputs five or six results nodes (the `+Univariate` node is optional and only appears if the user specifies the `showcurves` option in the `wtmodel` statement) in the results pane (vertically oriented pane, left of main window). These results nodes are identified as follows:

```
+ Logistic
+ Means
+ Univariate
+ Logistic | GLM
+ Logistic | GLM
+ Print
```

These nodes are as follows:

`+ Logistic`
This is the logistic regression for the propensity score model (`wtmodel`) portion of doubly robust estimation.

`+ Means`
This node provides the mean, standard deviation, minimum and maximum of the estimated weights.

`+ Univariate`
This is the graph of the propensity score curves generated by the wtmodel if the `showcurves` option is designated. If the `common_support` option is designated, this graph will have lines indicating the region of common support, as well.

`+ Logistic | GLM`
`+ Logistic | GLM`
If the distribution (`dist=`) option of the `model` statement is set to `bin`, then nodes will be labeled "`Logistic`"; if the distribution option is set to normal (`dist=n`), then these nodes will be "`GLM`". The first of these two nodes is the results of a model among the unexposed (or untreated), weighted such that this represents the expected response had all subjects in the analysis population been unexposed; the second is the model among the exposed (or treated) weighted such that this represents the expected response had all subjects in the analysis population been exposed.

`+ Print`
This results node produces output in the following format:

| Obs | totalobs | usedobs | dr1 | dr0 | deltadr | se |
|---|---|---|---|---|---|---|
| 1 | 100000 | 79292 | 0.034117 | .005546853 | 0.028570 | .002026204 |

Where:

`totalobs` is the number of observations in the specified dataset

`usedobs` is the number of observations actually used in the analysis. `usedobs` is always less than or equal to `totalobs`; `usedobs` is less than `totalobs` when there are missing values, and/or when the `common_support` option is used in the `wtmodel` statement.

`dr1` is an estimate of the average response that we would have observed if everyone in the population had been exposed or treated. This is estimated by a regression model adjusted for the covariates specified in the `model` statement in a pseudopopulation created by reweighting the exposed group using the propensity score from the `wtmodel` statement.

In the case of a continuous outcome, this is the mean value for that continuous variable (such as blood pressure, cholesterol, weight, etc). In the case of a dichotomous outcome, this is the average risk of the outcome. This is the expected mean response in subjects rather than the expected value for an average subject; these two values are the same in a linear model but not so for a logistic regression model which may lead to discrepancies.

`dr1` is Term 1 in Equation (9) in Lunceford and Davidian (2004).

`dr0` is an estimate of the average response that we would observe if everyone in the population had not been exposed (or not received the treatment). This is estimated by a regression model adjusted for the covariates specified in the `model` statement in a pseudopopulation created by reweighting the unexposed group using the propensity score from the `wtmodel` statement. (See `dr1`.)

`dr0` is Term 2 in Equation (9) in Lunceford and Davidian (2004).

`deltadr` is the difference (`dr1 – dr0`), and is equivalent to Equation (9) in Lunceford and Davidian (2004). In the case of a continuous outcome, this is the mean difference due to treatment or exposure. In the case of a dichotomous outcome, this is the difference in the average predicted probability of the outcome, comparing the response in the pseudopopulation as if everyone had been unexposed (or untreated) to the response in the pseudopopulation as if everyone had been exposed (or treated).

`se` is the standard error associated with the measure `deltadr` based on the sandwich estimator, (Lunceford and Davidian, 2004, Equation (21)).

## EXAMPLE 1

This is an example that simulates a dataset with a continuous response (`response1`), a dichotomous exposure (`tx`), and three covariates (`x1 x2` and `x3`) with a total sample size of n=4000 in which the outcome is unrelated to the exposure. Therefore, we would expect the difference in the mean `response1` under the two exposure conditions (`tx=1` and `tx=0`) to be not significantly different than 0.

```
data test;
 do i=1 to 4000;
    x1=rannor(101);
    x2=rannor(202);
    x3=ranuni(303)<.3;
    tx=ranuni(404)<.5;
    response1=rannor(505)+ x1 + x2;
    output;
 end;
run;
```

We run the doubly robust model on these simulated data using the following code:

```
title 'Example 1';
%dr(%str(options data=test descending;
       wtmodel tx = x1 x2 x3 / method=dr dist=bin ;
       model response1 = x1 x2 x3 / dist=n;));
```

This produces the output under `+ Print: Example 1`:

| Obs | totalobs | usedobs | dr0 | dr1 | deltadr | se |
|---|---|---|---|---|---|---|
| 1 | 4000 | 4000 | 0.039756 | 0.033922 | -.005833368 | 0.031623 |

This output can be interpreted as follows:

If all subjects in this cohort had been exposed (tx=1), the average outcome would have been 0.0339.
If all subjects in this cohort had been unexposed (tx=0), the average outcome would have been 0.0398.

`deltadr` is therefore the difference between the expected outcomes from the pseudo-populations of “all unexposed” compared to “all exposed”.

## EXAMPLE 2

The next example represents an analysis of simulated data where the exposure of interested is statin use (`statin`) and the outcome of interest is a continuous cardiovascular disease score (`rmi3a`). (To run this example, download the study dataset from http://www.harryguess.unc.edu and create a libname for ‘sampledata’ that points to the appropriate folder on your computer.) Both the weight (propensity score) model and the regression model are specified correctly in this first version of the analysis:

```
title 'Example 2';
%dr(%str(options data=sampledata.study descending;
       wtmodel statin=hs smk hxcvd black bmi age income chol exer
              / method=dr dist=bin showcurves common_support=.99;
      model rmi3a=hs female smk hxcvd bmi bmi2 age age2 chol exer /dist=n; ) );
```

The `showcurves` option produces a histogram that compares the distributions of the propensity score for the two levels of exposure with a nonparametric smoothed curve overlayed (See Figure 1). The `common_support` option indicates that the outcome regression models should ‘trim’ off observations that lie at the extreme ends of the propensity score distribution. Using `common_support=0.99`, the regression models are limited to those observations with a propensity score between the 0.5th percentile and the 99.5th percentile. The vertical dashed lines on Figure 1 indicate the boundaries for this portion of the data. In the final results shown below, the number of observations used (`usedobs`) is less than the total number of observations in the dataset (`totalobs`) because we specified the `common_support` option.

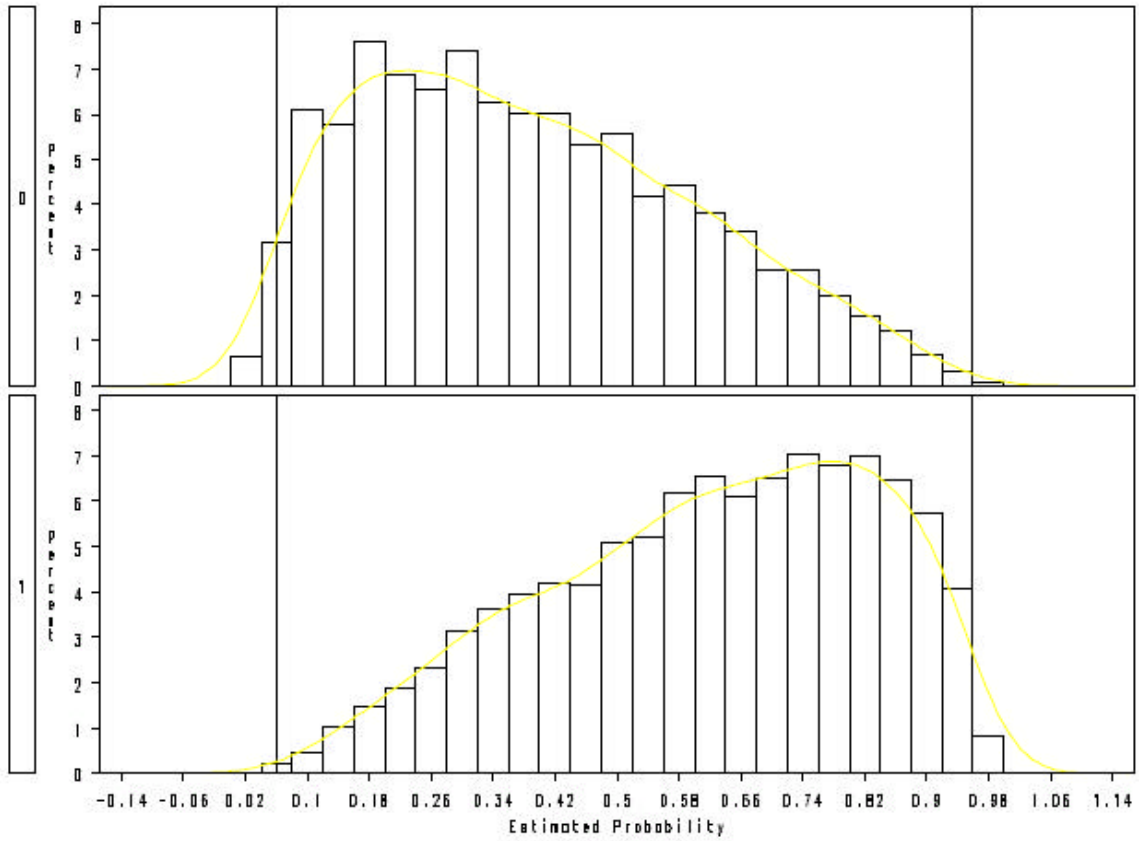


**Figure 1.** Estimated propensity score distributions stratified by exposure with nonparametric smoothed curve.

The doubly robust estimate of the average treatment effect when we have specified both models correctly is a difference of -1.10.

| Obs | totalobs | usedobs | dr0 | dr1 | deltadr | se |
|---|---|---|---|---|---|---|
| 1 | 10000 | 9852 | -8.83820 | -9.94092 | -1.10273 | 0.024158 |

If we alter the set of covariates included in the models to intentionally misspecify the weight or regression model by removing the covariate chol, we find that the doubly-robust estimator is unbiased when only one of the models is misspecified but is no longer unbiased when both models are misspecified.

**Table 1.** DR estimates from correctly and incorrectly specified models.

| | Model Specification | | | |
|---|---|---|---|---|
| | Weights | Outcome | DR Estimate | SE |
| Crude | | | 1.87 | 0.089 |
| DR | + | + | -1.10 | 0.024 |
| | + | - | -1.07 | 0.070 |
| | - | + | -1.09 | 0.022 |
| | - | - | 0.40 | 0.049 |

## EXAMPLE 3

The final example represents an analysis of simulated data where the exposure of interested is statin use (`statin`) and the outcome of interest is a dichotomous variable indicating whether or not the subject experienced a myocardial infarction within the follow-up period (`mi3`). (To run this example, download the study dataset from http://www.harryguess.unc.edu and create a libname for 'sampledata' that points to the appropriate folder on your computer.) Both the weight (propensity score) model and the regression model are specified correctly:

```
title 'Example 3';
%dr(%str(options data=sampledata.study descending;
       wtmodel statin=hs smk hxcvd black bmi age income chol exer
              / method=dr dist=bin showcurves;
      model mi3=hs female smk hxcvd bmi bmi2 age age2 chol exer /dist=bin; ) );
```

The doubly robust estimate of the average treatment effect when we have specified both models correctly is a risk difference of -0.014.

| Obs | totalobs | usedobs | dr0 | dr1 | deltadr | se |
|---|---|---|---|---|---|---|
| 1 | 10000 | 10000 | 0.039522 | 0.025832 | -0.013689 | .003676795 |

## CONSIDERATIONS

Development of the macro is ongoing, but at this time there are several limitations that users should be aware of. The exposure variable must be binary and coded as 0/1. The outcome may be binary (logistic regression) or continuous (linear regression).

Although we have intentionally designed the macro so that it behaves much like typical SAS procedures to improve its usability, there are some SAS conventions that are not currently implemented. Specifically, variables for interaction terms and higher order terms must be created in a data step – not within in the model statements. The class statement is also not recognized at this time and therefore all categorical variables should be coded using indicator variables.

Standard errors based on the sandwich estimator are known to be too conservative in small sample sizes. In this circumstance, bootstrapping provides a more reliable estimate of the standard error. We expect to implement this in a future release. As of version 0.90, the macro only calculates the doubly robust estimate for the *difference* between the average response or risk. In a future release, we will provide the relative risk and the odds ratio with the associated standard errors. In the meantime, one can manually calculate the relative effect measures from dr1 and dr0 but obtaining an appropriate standard error will require bootstrapping.

Please note that while University researchers created this program in good faith and have used it repeatedly, the University has not rigorously tested the tool as a commercial software provider would. The University welcomes information about any problems arising from use of the program.

## DOWNLOADING & INSTALLING

The DR macro can be downloaded from the Resources section of http://www.harryguess.unc.edu along with the sample data used for examples 2 and 3. Please review the Readme.pdf file first for important information regarding installation.

## CONCLUSION

Observational studies rely on data from non-randomized patients who have used the agent(s) of interest in the course of normal clinical care. Because patients and their health care providers decide whether and which agent to use, differing response rates and adverse event rates may reflect differences between the groups of patients rather than differences between the agents themselves. In the presence of confounding, an unbiased estimate of the effect of treatment can still be obtained if all of the factors that affect prognosis have been measured accurately. In order to obtain an unbiased estimate of the treatment effect, these confounding factors must be incorporated into a statistical model that represents the true relationship between each factor and the outcome of interest. Doubly robust estimation methods – which provide the analyst with two chances to correctly specify the true relationships among covariates in the data - are potentially valuable in studies of comparative effectiveness and other epidemiologic studies, but no standard software packages or user programs have previously been available to use this method. Our hope is that the availability of a SAS® macro will facilitate greater use of this method in the field of epidemiology broadly.

## REFERENCES

Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005 Dec;61(4):962- 73.

Davidian M. 2005. *Double Robustness in Estimation of Causal Treatment Effects*. Presentation to EPID 369: Modeling Causal Effects class, University of North Carolina at Chapel Hill, March 4, 2005.

Greenland S. 2004. An overview of methods for causal inference from observational studies. In: (Gelman A & Meng X-L, eds) Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives. John Wiley & Sons, 3-13.

Little RJ, Rubin DB. 2000. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health*, 21:121-45.

Lunceford JK, Davidian M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 23:2937-60.

Robins JM. 1998. Marginal structural models. In: 1997 *Proceedings of the Section on Bayesian Statistical Science.* Alexandria, VA: American Statistical Association, 1-10.

Robins JM. 1998a. Correction for non-compliance in equivalence trials. *Stat Med* 17:269-302.

Robins JM. 1999. Association, Causation, and Marginal Structural Models. *Synthese* 121, 151-179.

Robins JM. 1999a. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran E, Berry D, eds. *Statistical models in epidemiology: the environment and clinical trials*. New York, NY: Springer Verlag, 95-134.

Robins JM. 2000. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* 1999; 6-10.

Robins JM, Hernan MA, Brumback B. 2000. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 11(5), 550-560.

Robins J, Rotnitzky A, Zhao LP. 1994. Estimation of regression coefficients when some of the regressors are not always observed. *Journal of the American Statistical Association* 89:846-866.

Stefanski LA, Boos DD. The calculus of M-estimation. The American Statistician 2002; 56:29–38.

van der Laan MJ, Robins J. 2003. Unified methods for censored and longitudinal data and causality. Springer Verlag. New York.

## ACKNOWLEDGMENTS

The work was supported by AHRQ grant number 3U18HS010397-07S1. We would like to acknowledge the valuable contributions of Harry Guess, MD, PhD (deceased) to this project. We would also like to thank our colleagues at Duke University and students at the University of North Carolina who provided feedback on early versions of the macro.

## CONTACT INFORMATION

Michele Jonsson Funk, PhD
Department of Epidemiology
University of North Carolina
Chapel Hill, NC 27599-7521
mfunk@unc.edu
919-966-8431 (phone)
919-843-3120 (fax)

*This technical paper (#189-2007) was first presented at the SAS Global Forum Conference, April 18, 2007 in Orlando, Florida. This version of the paper includes updates and corrections not available at the time of the original technical paper's publication.*