# Hays, R. D., & Spritzer, K. L. (2013, November). REcycled SAS® PrEdiCTions (RESPECT).

**Paper presented at the Society for Computers in Psychology meeting, Toronto, Canada**

**Ron D. Hays**
UCLA Department of Medicine

**Karen L. Spritzer**
UCLA Department of Medicine

**Abstract**

Recycled predictions are used to understand the marginal effect of independent variables on a dependent variable. They are obtained from regression models by averaging predicted scores on the dependent variable after fixing the value of one independent variable (e.g., setting gender to be "female") and using observed values on the remaining independent variables in the sample. This paper presents a **SAS** macro, **RE**cycled **S**AS **P**r**E**di**CT**ions (RESPECT), that provides estimates from PROC REG and PROC GLM, contrasts among the recycled variables, and GLM least squares means (i.e., recycled predictions).

*Keywords*: **SAS®**, **STATA®**, recycled predictions.

## 1. Introduction

Recycled predictions are used to understand the marginal effect of independent variables on a dependent variable (Basu and Rathouz 2005; Graubard and Korn 1999). For example, a recent study examined the associations of 10 types of cancer (melanoma, endometrial cancer, colorectal cancer, female breast cancer, prostate cancer, bladder cancer, non-Hodgkin lymphoma, kidney cancer, non-small cell lung cancer, and other cancer) and 13 non-cancer chronic conditions (myocardial infarction/heart attack, hypertension, angina/coronary artery disease, congestive heart failure, other heart disease, diabetes, arthritis of the hand, inflammatory bowel disease, sciatica, stroke, chronic obstructive pulmonary disease/asthma, arthritis of the hip, depressive symptoms) with health-related quality of life in a sample of 126,366 Medicare beneficiaries (Hays *et al.* 2013). The authors used recycled predictions to show the unique associations of each of the 23 chronic conditions on the dependent variable (SF-36 preference-based score).

## 2. Methods

Recycled predictions are obtained from regression models by averaging predicted scores on the dependent variable after fixing the value of one independent variable (e.g., setting gender to be ”female”) and using observed values on the remaining independent variables in the sample. In the Hays *et al.* (2013) study, an ordinary least square regression model was run with dummy variables for the 23 conditions of interest, controlling for education, gender, marital status, age, race/ethnicity, income, whether a proxy completed the survey, and mode of administration (mail versus telephone). The recycled predictions provided adjusted means for each cancer and non-cancer condition. Two variants of the predicted scores were obtained. In one approach, all independent variables other than the condition being predicted were fixed at the sample means. In a second approach, the 22 conditions other than the one being predicted were fixed at zero and the remaining independent variables were fixed at their means. The advantage of the first approach is that adjusted scores correspond to the overall sample mean on the dependent variable. However, this approach can produce differences between the rank ordering of the recycled predictions and the regression beta weights. This issue does not occur when using the second approach, but the predicted scores are higher than the first approach and do not correspond to the dependent variable sample mean.

## 3. The Code

The reg procedure with ”predict” in **STATA** (StataCorp 2011) can be used to obtain recycled predictions. The example code below illustrates how to obtain a recycled prediction using the first approach mentioned above (fixing all independent variables other than the one for which recycled predictions are being done at their means) for two levels (0, 1) of an independent variable (iv1) in the case where a dependent variable (dv1) is regressed on three independent variables (iv1, iv2 and iv3):

```
. reg dv1 i.iv1 iv2 iv3
. replace iv1=0
. predict dv0
. replace iv1=1
. predict dv1
. gen dvdif=dv1-dv0
. tabstat dv0 dv1 dvdif
```

Equivalently and more efficiently, one can use the following alternative code:

```
. reg dv1 i.iv1 iv2 iv3
. margins iv1, atmeans
```

Li and Mahendra (2010) presented a **SAS**®(SAS Institute 2010) macro to obtain recycled predictions for logistic regression. **SAS** users can obtain recycled predictions for ordinary least squares regression models using **SAS**-callable **SUDAAN** (Research Triangle Institute 2001) code (Bieler *et al.* 2010): `http://www.rti.org/sudaan/pdf_files/110Example/REGRESS%20Example%201.pdf`.

We provide a **SAS** macro to obtain recycled predictions for ordinary least squares regression that does not require purchase of **SUDAAN**®software (Table 1). The output from the macro includes OLS estimates from PROC REG and PROC GLM, contrasts among the recycled variables, and GLM least squares means (i.e., recycled predictions). Table 2 provides an abbreviated output of the RESPECT macro using the Hays *et al.* (2013) example described above using the first recycled predictions approach that fixes all independent variables other than the one being predicted for at their means. The model has 43 degrees of freedom in the numerator and the adjusted R-squared is 0.39. The abbreviated output shows that the recycled mean for depressive symptoms (DEPMCH) is 0.6145 and for endometrial cancer (CORPUS) is 0.7245. Both of these recycled predictions appear in Table 2 of the published article (Hays *et al.* 2013).

## 4. Summary

The RESPECT macro provides recycled predictions for ordinary least square regression models run with the **SAS** programming software. These predictions are useful for understanding the marginal effects of independent variables on dependent variables in multivariate models.

## 5. Author Note

This paper was presented at the annual meeting of the Society for Computers in Psychology, November 14, 2013, Toronto, Canada. Ron D. Hays was supported by NIH/NIA Grants P30-AG028748 and P30-AG021684, and NCMHD Grant 2P20MD000182. Karen L. Spritzer was supported by discretionary funds of Ron D. Hays.



## 6. Table 1

```
options ls=132 ps=54 nonumber nocenter dquote nofmterr;
*-----------------------------------------------------;
* place directory where your dataset is located within double quote marks;
libname lib "c:\mylib";

* dataset name;
%let ds=mydata;

* outcome variable;
%let outcome = sf6d36v1;
```

```
* list recycling variables;
%let recycle = depmch corpus MELANOMA  COLOR BREAST PROSTATE
BLADDER NHLYMP KIDNEY OTHERC LUNG MI HYPER ANGINA OTHHRT DIAB HAND CHF
GI SCIAT STROKE COPD HIP;

* list any additional independent (casemix) variables;
%let adjust = educ male married widowed mail age proxy hisp black asian
ai other missing INC1 INC2 INC3 INC4 INC5 INC6 INC7;

/* if no additional independent variables then leave blank, as in..
   i.e., uncomment this version and use it:
%let adjust = ;
*/

/*********************************************/
/*** DO NOT ALTER ANYTHING BELOW THIS LINE ***/
/*********************************************/

/* macro used to obtain recycled predictions */

%macro respect();

TITLE "Present PROC REG/STB estimates"; run;
proc reg data=lib.&ds;
model &outcome = &recycle &adjust /stb;
run;

/** Approach #1 - fix all indpendent variables other than the condition
  being predicted at the sample means **/

TITLE "Approach #1: GLM, estimates, and LSmeans follow.."; run;

* get number of recycling variables;
%let num_recv=%sysfunc(countw(&recycle));

proc glm data=lib.&ds;
class &recycle;
model &outcome = &recycle &adjust;
  %do i = 1 %to &num_recv;
      %do j = 1 %to &num_recv;
/* use this form instead to cut out redundant terms:  %do j = 1 %to &num_recv -1;  */
      %if &i ne &j %then %do;
        %let nv1 = %scan(&recycle, &i);        /* nv1: 1st variable to estimate */
        %let nv2 = %scan(&recycle, &j);        /* nv2: 2nd variable to estimate */
        estimate "&nv1 vs &nv2" &nv1 -1 1 &nv2 1 -1 ; /* estimate cross-products */
      %end;
      %end;
```

```
  estimate "&nv1" &nv1 -1 1;            /* get point estimate for NV1 */
  %end;
lsmeans &recycle /om stderr;
run;
quit;
/*****************************************************************/

/** Approach #2: pick one condition and fix all other recycling vars to
   zero (keep adjusting vars at mean) **/

/* get 0's for use in building lsmeans statement in GLM (one less than total
        number of recycling vars */

%let zeros = ;
%do k = 1 %to %eval(&num_recv -1);
%let zeros = &zeros 0 ;
%end;

/** build a GLM for each recycling variable **/
 %do ii = 1 %to &num_recv;

        /** initialize item you're keying in on (keyvar) and drop it from list
               of (remain)ing recycling vars for use in building GLM **/
%let keyvar =;
%let remain =;

%let keyvar = %scan(&recycle, &ii);
%do jj = 1 %to &num_recv;
        %if &ii ne &jj %then %do;
        %let remain = &remain %scan(&recycle, &jj);
     %end;
%end;

TITLE1 "Approach #2: pick one condition (&keyvar) and fix all other";
TITLE2 "recycling vars to zero (keep adjusting vars at mean)"; run;

proc glm data=lib.&ds;
class &keyvar;
model &outcome =  &recycle  &adjust;
lsmeans &keyvar
 /at ( &remain ) =
( &zeros ) out=ls_remain;
run;

%end;
%mend;
%respect();
```

# 7. Table 2

```
Present PROC REG/STB estimates

The REG Procedure
Model: MODEL1
Dependent Variable: sf6d36v1

Number of Observations Read      126366
Number of Observations Used      126366


                             Analysis of Variance

                                       Sum of           Mean
Source                   DF           Squares         Square     F Value     Pr > F

Model                    43         913.48555       21.24385     1860.32     <.0001
Error                126322        1442.52798        0.01142
Corrected Total      126365        2356.01353


Root MSE              0.10686    R-Square      0.3877
Dependent Mean        0.73000    Adj R-Sq      0.3875
Coeff Var            14.63853


                              Parameter Estimates

                         Parameter       Standard                                  Standardized
Variable     DF           Estimate          Error     t Value     Pr > |t|            Estimate

Intercept    1             0.80102        0.00375      213.83       <.0001                   0
DEPMCH       1            -0.13081     0.00097812     -133.73       <.0001            -0.30819
CORPUS       1            -0.00553        0.00359       -1.54       0.1227            -0.00341
...


Approach #1: GLM, estimates, and LSmeans follow..

The GLM Procedure

Number of Observations Read      126366
Number of Observations Used      126366
```

*Dependent Variable: sf6d36v1*

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 43 | 913.485550 | 21.243850 | 1860.32 | <.0001 |
| Error | 126322 | 1442.527978 | 0.011419 | | |
| Corrected Total | 126365 | 2356.013528 | | | |

| R-Square | Coeff Var | Root MSE | sf6d36v1 Mean |
|---|---|---|---|
| 0.387725 | 14.63853 | 0.106862 | 0.730004 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| DEPMCH | 1 | 431.5488876 | 431.5488876 | 37790.7 | <.0001 |
| CORPUS | 1 | 0.3050781 | 0.3050781 | 26.72 | <.0001 |

...

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| depmch vs corpus | -0.12527341 | 0.00371339 | -33.74 | <.0001 |

...

*Approach #1: GLM, estimates, and LSmeans follow..*

*The GLM Procedure*
*Least Squares Means*

| DEPMCH | sf6d36v1 LSMEAN | Standard Error | Pr > \|t\| |
|---|---|---|---|
| 0 | 0.74533962 | 0.00032174 | <.0001 |
| 1 | 0.61453125 | 0.00091428 | <.0001 |

| CORPUS | sf6d36v1 LSMEAN | Standard Error | Pr > \|t\| |
|---|---|---|---|
| 0 | 0.73004331 | 0.00030170 | <.0001 |
| 1 | 0.72450836 | 0.00357304 | <.0001 |

...

## References

Basu A, Rathouz PJ (2005). "Estimating Marginal and Incremental Effects on Health Outcomes Using Flexible Link and Variance Function Models." *Biostatistics*, **6**(1), 93–109.

Bieler GS, Brown GG, Williams RL, Brogan DJ (2010). "Estimating Model-adjusted Risk Differences and Risk Ratios from Complex Survey Data." *American Journal of Epidemiology*, **171**(5), 618–623.

Graubard BI, Korn BL (1999). "Predictive Margins with Survey Data." *Biometrics*, **55**(2), 652–9.

Hays RD, Reeve BB, Smith AW, Clauser SB (2013). "Associations of Cancer and Other Chronic Conditions with SF-6D Preference-Based Scores in Medicare Beneficiaries." *Quality of Life Research*.

Li Z, Mahendra G (2010). "Using 'Recycled Predictions' for Computing Marginal Effects." In *Proceedings of the **SAS**® Global Forum 2010 Conference*. **SAS** Institute, Inc., Cary, NC, USA. URL http://support.sas.com/resources/papers/proceedings10/272-2010.pdf.

Research Triangle Institute (2001). ***SUDAAN** User's Manual: Release 8.0*. Triangle Park, NC. URL http://www.rti.org/sudaan/.

SAS Institute (2010). ***SAS**® 9.3 TS Level 1M1*. Cary, NC, USA. URL http://www.sas.com/.

StataCorp (2011). *Stata Statistical Software: Release 12*. College Station, TX. URL http://www.stata.com/.

**Affiliation:**

Ron D. Hays, Ph.D.
UCLA Department of Medicine,
Division of General Internal Medicine and Health Services Research
911 Broxton Avenue
Los Angeles, CA 90024 U.S.A.
Telephone: +1/310/794-2294
Fax: +1/310/794-0732
E-mail: drhays@ucla.edu
URL: http://gim.med.ucla.edu/FacultyPages/Hays/