

# A MACRO PROCEDURE IN SAS® SOFTWARE TO SUMMARIZE OUTPUT OF REGRESSION MODELS

Chih-Ping Chou, University of Southern California  
Eric Yu I. Wang, University of Southern California

## Abstract

A program, REGRPT, has been developed to summarize results obtained from several regression models using the same set of covariates in the REG procedure. This program combines some macros and SAS/IML with results from REG and MEAN procedures to report the regression weight and standard error of each covariate on each dependent variable as well as the  $R^2$  and adjusted  $R^2$  values for each model. It also allows users to specify a subset of covariates to be compared and a subset which are also in the models but not to be reported. Symbols are also provided for regression weights which are significant at different levels. The program definitely provides an efficient and error-free method to offer a more clear comparisons among several regression models.

## Introduction

Linear regression model has been one of the most widely used techniques in analyzing functional relationships among variables. A regression model can be expressed as:

$$Y = b_0 + X_1b_1 + X_2b_2 + \dots + X_kb_k + c \quad (1)$$

It is also an extremely popular method for prediction based on knowledge of a set of covariates.

In the area of epidemiology research, the impacts of a set of covariates on separate outcome variables usually need to be investigated. Equation (1) can be applied and evaluated for different Y's while the X's remain the same. The statistics of major concerns to evaluate the appropriateness of the models and the contribution of each covariate in regression models are the  $R^2$  value, adjusted  $R^2$  value, parameter estimates, and standard errors of regression weights. A summary report containing all these statistics will be very informative in investigating the effect of each covariate on each outcome. In general, various results from the regression models can be obtained from the REG, RSQUARE, and STEPWISE procedures in SAS. However, it is unfortunate that there is no straight forward way to output the estimated standard errors and it requires several procedures to obtain all the above information. For example, the information of  $R^2$  can not be output to a file from the REG procedure. The RSQUARE can output  $R^2$  or adjusted  $R^2$ , but not both. The STEPWISE does not even have the OUTEST statement to output file containing analysis results.

To summarize the results for comparisons among different models by looking through large amounts of printed output is very time consuming. This program, REGRPT, provides a more efficient and error-free approach by using a macro along with SAS/IML programming. Using the OUTEST in the REG procedure, one can easily obtain regression weights. Standard errors can be located from the diagonal of the covariance matrix of estimates. The  $R^2$  and adjusted  $R^2$  values can be computed as a function of the MSE and the variances of Y yielded from the MEAN procedure. The REGRPT program provides users an options to select a subset of covariates while ignoring other covariates in the models for comparisons. In order to be consistent with the REG procedure in handling missing data, this program evaluates each model separately. With some minor modification, the process can be more efficiently implemented with a slight modification if the missing pattern for all the outcome variables are the same. Finally, specially symbols are provided for the regression weights which are significant at some predetermined levels. The

summarized results are automatically put into a permanent file and ready to be incorporated into a report.

## Input Required for REGRPT

The REGRPT program is attached in Appendix 1. The program can be used with all the systems where SAS/IML and MACRO are acceptable, e.g. PC, MVS, CMS, UNIX, etc.

The program requires the following input section:

- (1). I/O DESTINATION section to provide the destination of the input and output files.
- (2). MODEL SPECIFICATION section to specify the set of covariates and dependent variables in the regression models and also the subset of covariates to be reported and the subset not to be reported.
- (3). DATA PROCESS section to input the data and conduct data transformation.

More information regarding the input required by the program can be found in the Appendix 1, where an example is provided.

## Example

The example provided in Appendix 1 analyzes the impacts of some demographical information on six measures of substance use. The six dependent variables are: cigarette monthly use (MNC), cigarette weekly use (WNC), alcohol monthly use (MNA), alcohol weekly use (WNA), marijuana monthly use (MNM), and marijuana weekly use (WNM). Six regression models, one for each substance use measure, are constructed along with the same set of independent variables. The covariates included in the models are RACE, SES, GROUP membership, TIME, and some interactions among RACE, SES, and GROUP, as well as square and cubic of TIME variables. However, only the first four covariates are to be compared across different models. The REGRPT.SAS program output the summarized results from all six models to REGRPT.OUT file (see Appendix 2). The results showed that only the effects of RACE is consistent with all six measures of substance use. SES has impact on cigarette and marijuana use but not on alcohol use. GROUP and TIME only have scattered impact on substance use. The final results can also be obtained by typing in corresponding information summarized from 12 pages of printed output.

## Discussion

The advantages of the REGRPT.SAS program over the conventional ways in summarizing the statistical results are obvious. This program provides a more efficient method in presenting the results. More importantly, this method can avoid human error often introduced through typing and during summarizing of the results. A possible shortcoming of this program is, that the output file can not directly be incorporated into a written report. For example, the name of the dependent variables is still not part of the original output in the REGRPT.OUT file. Thus, more editing is still needed to make it a complete table. This problem may be resolved if a newer version of the TABULATE procedure in SAS is developed.

```

*****
*
*          PROGRAM : REGRPT.SAS
*
*  PURPOSE:
*          THIS PROGRAM SUMMARIZES RESULTS
*          FROM REGRESSION MODELS. IT MAY BE
*          USED TO COMPARE SEVERAL MODELS
*          USING THE SAME SET OF COVARIATES.
*          THE SUMMARIZED RESULTS MAKE THE
*          COMPARISONS EASIER AMONG MODELS.
*
*  PROGRAMMER:
*          CHIH-PING CHOU & ERIC WANG
*
*  DATE CREATED: AUGUST 23, 1991.
*
*  DATE OF LAST REVISION: AUGUST 23, 1991.
*
*****;

*+++++*
* THE FOLLOWING TWO SECTIONS NEED TO BE REVISED
* BY USERS
*+++++*
* I. I/O DESTINATION (USER DEFINE SECTION)
*
* THE FOLLOWING TWO STATEMENTS ARE NEEDED TO
* SHOW:
*     1. LIBRARY WHERE DATA SET IS LOCATED,
*     2. OUTPUT FILE FOR THE REPORT.
*
*+++++*

LIBNAME IN  'u/chou/memol';
FILENAME OUT1 'u/chou/regprt.out';
OPTIONS PS=58 LS=132;

*+++++*
* II. MODEL SPECIFICATION (USER DEFINE SECTION)
*
* THE FOLLOWING SECTIONS WITH LET STATEMENTS ARE
* REQUIRED INPUT INFORMATION:
*
* 1. DEPVAR: NAME OF DEPENDENT VARIABLES.
* 2. COVRPT: NAME OF INDEPENDENT VARIABLES TO BE
*   REPORTED.
* 3. COVNRP: NAME OF INDEPENDENT VARIABLES NOT TO
*   BE REPORTED.
*+++++*

*** DEPVER IS FOR DEPENDENT VARIABLES ***;
%LET DEPVAR=MNC WNC MNA WNA MNM WNM;

*** COVRPT IS FOR THE COVARIATES TO BE REPORTED
***;
%LET COVRPT=TM GROUP RACE SES;

*** COVNRP IS FOR THE COVARIATES IN TH EQUATION ***

```

```

***          WHICH ARE NOT TO BE REPORTED          ***;
%LET COVNRP= TM2 TM3 ST RT SG RG;

*+++++*
* THE ABOVE TWO SECTIONS NEED TO BE REVISED BY
* USERS
*+++++*

DATA B;
PROC IML;
START MAIN;
V1={&DEPVAR};
V2={&COVRPT};
V3={&COVNRP};

NDEP=NCOL(V1);
NRPT=NCOL(V2);
NNRP=NCOL(V3);
NV=NDEP'NRPT'NNRP;
CNAME={'NDEP' 'NRPT' 'NNRP'};
PRINT NV [COLNAME=CNAME];

CREATE INFI FROM NV [COLNAME=CNAME];
APPEND FROM NV [COLNAME=CNAME];

FINISH;

RUN MAIN;

DATA INFO;
SET INFI;
CALL SYMPUT('NDEP',NDEP);
CALL SYMPUT('NRPT',NRPT);
CALL SYMPUT('NNRP',NNRP);
RUN;

%LET NDEP=&NDEP;
%LET NRPT=&NRPT;
%LET NNRP=&NNRP;

%MACRO OSET(NUMBER=);
%DO I = 1 %TO &NUMBER;
O&I
%END;
%MEND OSET;

*****
* PROVIDE INFORMATION OF INPUT AND OUTPUT FILES
*
* FOR MACRO REG
* EX: %REG(INPUT=A,OUTPUT=IN.REGLM6)
*****;

%MACRO REG(DVAR=,MODELN=);

PROC REG COVOUT DATA=A OUTEST=REGOUT ; /* REGOUT
*/
MODEL &DVAR = %STR(&COVRPT)
%STR(&COVNRP)/SEQB;

DATA O&MODELN;
SET REGOUT;
IF N_ < (&NRPT+3);
KEEP _RMSE_ INTERCEP %STR(&COVRPT);
%MEND REG;

```

```

*
* PROVIDE THE INFORMATION OF INPUT AND OUTPUT FILES
*
* EX:  %REPORT(INPUT=IN.REGM6,OUTPUT=OUT)
*
*;

%MACRO REPORT(OUTPUT=);

DATA B;
SET A;

ARRAY X %STR(&COVRPT) %STR(&COVNRP);
ARRAY Y %STR(&DEPVAR);
ARRAY V V1-V&NDEP;
ARRAY NT NT1-NT&NDEP;

DO OVER X;
  IF X=. THEN DELETE;
END;

PROC MEANS N VAR;
VAR %STR(&DEPVAR);
OUTPUT OUT=MEANOUT
      N=NT1-NT&NDEP VAR=V1-V&NDEP;

DATA C;
SET %OSET(NUMBER=&NDEP); /* SAS MACRO P.48 */

DATA B;

PROC IML;

START MAIN;
NDEP={&NDEP};
NRPT={&NRPT}+1;
NNRP={&NNRP};
NCOV=NRPT+NNRP;

USE C;
READ ALL INTO REG;

USE MEANOUT;
READ ALL INTO TVN;

RNAME='INTERCEP' '{&COVRPT}' 'RSQUARE' 'RSQRADJ'
      'N_COV' 'TOTAL_N';
NC=NCOL(TVN);
TVN=TVN[,3:NC];
N=TVN[,1:NDEP];
V=TVN[,NDEP+1:(2*NDEP)];
NC=NCOL(REG);
RMSE=REG[,1];
REG =REG[,2:NC];
PRINT N V;
PRINT REG;

/* COLNAME SPECIFY THE VARIABLE NAMES TO BE SENT TO
DATA REPORT */
COLNAME={ 'EST1' 'STE1' 'TRATIO1' 'P1'
          'EST2' 'STE2' 'TRATIO2' 'P2'
          'EST3' 'STE3' 'TRATIO3' 'P3'
          'EST4' 'STE4' 'TRATIO4' 'P4'
          'EST5' 'STE5' 'TRATIO5' 'P5'

```

```

          'EST6' 'STE6' 'TRATIO6' 'P6'};

JRPT=J(NRPT,1,1);
DO I = 1 TO NDEP;
  I1=(I-1)*(NRPT+1)+1;
  IIP=I1+1;
  E=REG[I1,];
  E=E';
  IIX=I1+NRPT;
  T=REG[IIP:IIX,1:NRPT];
  S=VECDIAG(T);
  S=SQRT(S);
  T=E/S;
  NN=N[1,I];
  DF=NN-NRPT-NNRP;
  P=PROBT(T,DF);
* IF P>(0.5*JRPT) THEN P=JRPT-P;
* P=2*P;
  T1=(E'S'T'P);
  SGE=RMSE[I1,1]**2;
  VY=V[1,I];
  RSQADJ=1-SGE/VY;
  T=(1-RSQADJ);
  RSQ=1-T*(NN-NCOV)/(NN-1);
  TT=J(4,4,-99); /* TT IS USED FOR RSQ, RSQADJ, AND N */
  TT[1,1]=RSQ;
  TT[2,1]=RSQADJ;
  TT[3,1]=NCOV-1;
  TT[4,1]=NN;
  T1=T1/TT;
  RST=RST'T1;
END;

PRINT RST;

CREATE A FROM RST [ROWNAME=RNAME
COLNAME=COLNAME];
APPEND FROM RST [ROWNAME=RNAME];

FINISH;

RUN MAIN;

DATA REPORT;
SET A;
FILE OUT1;
ARRAY EST EST1-EST6;
ARRAY STE STE1-STE6;
ARRAY TRATIO TRATIO1-TRATIO6;
ARRAY PROB P1-P6;
ARRAY SG $ SG1-SG6;

DO OVER STE;
IF TRATIO=-99 THEN TRATIO=.;
IF STE =-99 THEN STE =.;
IF PROB =-99 THEN PROB =.;
SG=' ';
IF PROB>0.5 THEN PROB=1-PROB;
PROB=2*PROB;
IF PROB<0.10 THEN SG='+';
IF PROB<0.05 THEN SG='*';
IF PROB<0.01 THEN SG='$';
IF PROB=. THEN SG='';
END;

```

