Paper 269

Using SAS® to Conduct Pilot Studies: An Instructors Guide

Sean W. Mulvenon, University of Arkansas, Fayetteville, AR
Ronna C. Turner, University of Arkansas, Fayetteville, AR

ABSTRACT

An important component in the instruction of educational research methods and statistics courses is the practical benefit of completing pilot studies. The use of the SAS® programming language has become an important component in teaching students how to conduct preliminary pilot studies before conducting research projects that require the collection of data. The outcome has been students that are more aware of how to utilize the SAS® programming language and appreciate the utility of replicating published work to generate research ideas. The programs in this paper have been designed to work on any DOS operating system which can operate the SAS® package and for users with intermediate expertise in using the SAS® programming language. The SAS® packages utilized were BASE, SAS/STAT, and IML.

INTRODUCTION

An important component in the instruction of educational research methods and statistics courses is the practical benefit of completing pilot studies. A pilot study can be used as the basis for conducting larger more comprehensive studies by providing a method for checking the reliability and validity of instruments or methodologies proposed for a study. In addition, a pilot study can help a researcher determine appropriate sample size and design components for achieving a desired level of power for a given effect size. Thus, the introduction of this procedure is important from an instructional viewpoint for the aspiring graduate student who plans to conduct educational research. It is important to encourage potential researchers to make a priori determinations of necessary sample sizes and appropriate research models for conducting proposed analytical studies.

The purpose of this paper is to demonstrate some instructional methods provided to graduate students at the University of Arkansas in order to facilitate the use of pilot studies using previously published data. The methods proposed include a variety of SAS® programming examples and can be used by most programmers.

DATA COLLECTION

As with most research projects, the collection of data can be very costly and time consuming. However, before proceeding with numerous surveys or questionnaires to conduct a pilot study, an alternative approach is suggested. Most studies, even pilot studies, require an extensive review of the literature of the intended area of research. The review of research frequently includes empirical studies containing data, typically descriptive data, that can be beneficial for use in pilot studies.

The type of data presented dictates the type of pilot study that can be completed, however, the most common data provided are means, standard deviations, and correlations. The use of these descriptive statistics is all that is required to complete replications of the work provided in these studies. For example, consider a researcher who is interested in investigating suicide potential in teenage females. After completing a review of the literature, the researcher has several examples of the studies completed on this topic and the variables most often associated with suicide potential for females. The researcher uses these results to develop his/her own theories, typically focusing on the variables already presented in the literature. The goal of the pilot study is to use the descriptive statistics from previously conducted studies to test the feasibility of all or portions of one's own hypotheses. Thus, if the descriptive statistics are available, the researcher could complete some preliminary pilot work using the data provided in these studies.

MULTIPLE REGRESSION STUDIES

A common methodology that is easy to replicate and

"use for further research" is a multiple regression procedure. The multiple regression procedure can be employed if a study provides the correlations and the sample size. It is beneficial if means and standard deviations are also reported to help provide unstandardized regression coefficients, but it is not mandatory. The following program demonstrates a method for completing a multiple regression pilot study.

```
 /*   Sample Program I:  A Multiple Regression   */

Data one(type=corr);
infile cards missover;
input _type_ $ _name_ $ x1 x2 x3 x4 x5;
cards;
n . 100 100 100 100 100

/* Optional Information */

mean . 45 22 38 71 65
std  . 7 8 5 12 18

corr x1 1.00
corr x2  .35 1.00
corr x3  .40  .15 1.00
corr x4  .52  .37  .31 1.00
corr x5  .48  .40  .50  .46 1.00
;

proc reg;
 model x1 = x2 x3 x4 x5 / scorr1 scorr2 influence;
 model x5 = x3 x2 x4 x1 / scorr1 scorr2 influence;
run;

            /*    End of Sample Program I    */
```

The output produces adjusted r-square which is an estimate of the projected replication in a secondary study.

KEY ELEMENTS OF MULTIPLE REGRESSION

Important aspects of this program include the sample size (for determination of the statistical significance, the default in SAS® is n = 10,000) and the correlation matrix. The command missover used with the infile statement requires only the diagonal and lower half of the correlation matrix be reported. There are a number of instances where a researcher may want to either test a different model than that used in a previous study or attempt to control for extraneous variability that was not accounted for previously, such as using a suppressor variable. The multiple model statements allow the researcher to evaluate different combinations of the variables to be utilized for exploration of possible modifications to the theories provided in the study being replicated.

MULTIVARIATE ANALYSIS OF VARIANCE

A multivariate analysis of variance (MANOVA) is another common methodology used in educational research. The use of MANOVA, when raw data is not provided, requires the reporting of group means, standard deviations, and correlation matrices. The following program provides an example of a procedure to replicate results from a study or to model results using a multivariate approach. In this program, the means, standard deviations, and correlations have been entered using IML for the statistical analysis. This study is an example of research from a dissertation (Rose, 1998) investigating the differences in OLSAT, Stanford-9 Reading, and Stanford-9 Mathematics scores comparing students who did and did not attend summer school.

```
/*****************************************/

Sample Program II: A Multivariate Analysis of
Variance

/*****************************************/

proc iml;
start manova;
n1= 383; n2= 54; N= n1 + n2;

/* p is the number of dependent variables */
p = 3;
mean1= {98.33 53.31 46.08};
mean2= {85.69 25.44 27.48};
corr1= {1 .66 .68, .66 1 .69, .68 .69 1};
corr2= {1 .44 .59, .44 1 .66, .59 .66 1};
sd1= {16.11 29.22 28.79};
sd2= {12.13 18.53 20.26};

/* calculate sscp matrices */
```

```
ss1= (n1-1)*(diag(sd1)*corr1*diag(sd1));
ss2= (n2-1)*(diag(sd2)*corr2*diag(sd2));

/* pool matrices */

S= (ss1 + ss2)/ (n1 + n2 - 2);

/* invert matrix */

S_inv= inv(S);

/* calculate Hotellings T2 */

T_2=    (n1*n2)/(n1 + n2) * (mean1 - mean2) *
        S_inv * (mean1 - mean2)`;

/* calculate Mahalonobis Distance (D2) */

D2= (mean1 - mean2)* S_inv *(mean1 - mean2)`;
F= (n1 + n2 - p - 1)/((n1 + n2 - 2)*p) * T_2;

/*****************************************/
```

The next section calculates the Univariate F-tests for each dependent variable and the multivariate effect sizes

```
/*****************************************/

meanvec= (mean1- mean2);
s_diag=diag(s);
s1_diag= inv(s_diag);
F_var= meanvec * s1_diag;
F_univ= (n1*n2)/(n1 + n2) * (meanvec[1,1] *
        F_var[1,1]//meanvec[1,2] * F_var[1,2]//
        meanvec[1,3]*F_var[1,3]);

df= N - 2;
p_f1= round(1-probf(f_univ[1,1], 1, df), .01);
   if p_f1 < .01 then p_f1= '< .05';
p_f2= round(1-probf(f_univ[2,1], 1, df), .01);
  if p_f2 < .01 then p_f2= '< .05';
p_f3= round(1-probf(f_univ[3,1], 1, df), .01);
  if p_f3 < .01 then p_f3= '< .05';
prob= p_f1//p_f2//p_f3;


/* calculate Retrospective and Prospective Power
Estimates */

sr_mse= inv(sqrt(diag(s)));
delta= round(t(meanvec*sr_mse), .01);
effect_p= {20 10 15};
delta_p= round(t(effect_p*sr_mse), .01);
ncp1= fnonct(delta[1,1], 1, df, .05);
fcrit1= finv(.95, 1, df);
ncp2= fnonct(delta[2,1], 1, df, .05);
fcrit2= finv(.95, 1, df);
ncp3= fnonct(delta[3,1], 1, df, .05);
fcrit3= finv(.95, 1, df);
ncp4= fnonct(delta_p[1,1], 1, df_p, .05);
fcrit4= finv(.95, 1, df_p);
ncp5= fnonct(delta_p[2,1], 1, df_p, .05);
fcrit5= finv(.95, 1, df_p);
ncp6= fnonct(delta_p[3,1], 1, df_p, .05);
fcrit6= finv(.95, 1, df_p);
power1= 1 - probf(fcrit1, 1, df, ncp1);
power2= 1 - probf(fcrit2, 1, df, ncp2);
power3= 1 - probf(fcrit3, 1, df, ncp3);
power= power1//power2//power3;
power_1= 1 - probf(fcrit4, 1, df_p, ncp4);
power_2= 1 - probf(fcrit5, 1, df_p, ncp5);
power_3= 1 - probf(fcrit6, 1, df_p, ncp6);
power_p= power_1//power_2//power_3;

title2 "MANOVA using SAS IML";
print '***   Multivariate F-test   ***';
print '          ' F '          ';
print ' ';
print '***   Univariate F-tests    ***';
names1={olsat reading math};
print '          'f_univ[rowname=names1] " prob '   ';
print ' ';
print '***   Retrospective and Prospective Power
          Estimates   ***';
print '          'power[rowname=names1] delta '  '
           power_p p_delta '        ';
finish manova;
run manova;

          /*    End of Sample Program II    */
```

KEY ELEMENTS OF MANOVA

Aside from the essential information from the descriptive statistics, it is important the researcher have a solid foundation in MANOVA. A review of the program provides the basis for this foundation, while the complexity of the program gives insight into the difficulty of MANOVA. The procedure

have a solid foundation in MANOVA. A review of the program provides the basis for this foundation, while the complexity of the program gives insight into the difficulty of MANOVA. The procedure allows the use of MANOVA when ANOVA's, regression, or other research designs were used in the previous studies. In addition, retrospective and prospective power estimates are computed for determining power of the model and as an aid for selecting sample size for the future study.

EXAMPLES USING RESEARCH ARTICLES

A classroom exercise currently used is the replication and evaluation of published research. The student is required to cite an article and provide the data. The following are examples of output. The first example is data from an article in The Journal *of Applied Psychology* (Adams, King, & King, 1996) investigating the relationship among job and life satisfaction using job and family background characteristics. The replications include the evaluation of an endogenous variable for a path analysis model and the predictive power of a job involvement variable when included in a model predicting life satisfaction.

Example 1

```
data one(type=corr);
infile cards missover;
input _type_ $ _name_ $ job-i family-i int-asst
      emot_sus work-int fam_int job-sat
      life-sat;

label
    job-i= job involvement
 family-i= family involvement
 int_asst= instrumental assistance
 emot_sus= emotional sustenance
 work_int= work interfering with family
  fam_int= family interfering with work
   job-sat= job satisfaction
 life-sat= life satisfaction;

cards;
n . 146 146 146 146 146 146 146 146
mean . 3.00 3.77 3.52 3.80 3.11 2.11 3.41 5.74
std . .72 .72 .77 .60 .81 .64 .80 .93
corr job-1    1.00
corr family-1 -.ll 1.00
corr int-asst -.03 .01 1.00
corr emot-sus -.04 .22 .63 1.00
corr work-int .28 .05 -.26 -.25 1.00
corr fam_int .10 .05 -.39 -.30 .30 1.00
corr job_sat .29 .03 .17 .21 -.24 -.14 1.00
corr life-sat -.11 .19 .28 .39 -.25 -.16 .29 1.00


proc reg;
 model life-sat = int-asst emot-sus work-int
                  job-sat / scorrl stb vif;
 title2 "replication of endogenous variable in
        path model analysis";
run;

proc reg;
 model life-sat = emot-sus job-sat job-i/ scorrl
                  scorr2 stb;
 title2 "revised model with job involvement as
        additional predictor";
run;
```

/* Statistical Output */

The SAS System
replication of endogenous variable in path model analysis

Model: MODEL1
Dependent Variable: LIFE-SAT life satisfaction

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	26.53644	6.63411	9.461	0.0001
Error	141	98.87406	0.70123		
C Total	145	125.41050			

Root MSE	0.83740	R-square	0.2116
Dep Mean	5.74000	Adj A-sq	0.1892
C.V.	14.58881		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for HO: Parameter=0
INTERCEP	1	3.516115	0.66733058	5.269
INT-ASST	1	0.029197	0.11742306	0.249
EMOT-SUS	1	0.471039	0.15122719	3.115
WORK-INT	1	-0.139429	0.09122679	-1.528
JOB-SAT	1	0.224278	0.09073188	2.472

Variable	DF	Prob > |T|	Standardized Estimate	Squared Semi-partial Corr Type I
INTERGEP	1	0.0001	0 .oooooooo	.
INT-ASST	1	0.8040	0.02417382	0.07840000
EMOT-SUS	1	0.0022	0.30389625	0.07585074
WORK-INT	1	0.1287	-0.12143825	0.02338085
JOB-SAT	1	0.0146	0.19292706	0.03416502

Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
INT-ASST	1	1.69041357
EMOT-SUS	1	1.70242120
WORK-INT	1	1.12906669
JOB-SAT	1	1.08944308

```
                    Variable
Variable   DF       Label

INTERCEP    1  Intercept
INT_ASST    1  instrumental  assistance
EMOT-SUS    1  emotional  sustenance
WORK-INT    1  work  interfering  with  family
JOB-SAT     1  job  satisfaction

                        The SAS System
   revised model with job involvement as additional predictor

Model: MODEL1
Dependent Variable: LIFE-SAT   life satisfaction

                        Analysis of Variance

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

Model         3     20.27470       9.42493       13.778     0.0001
Error       142     97.13572       0.66405
C Total     145    125.41050

    Root MSE          0.82706     R-square        0.2255
    Dep Mean          5.74000     Adj R-sq        0.2091
    C.V.             14.40899

                     Parameter  Estimates

                     Parameter       Standard     T for HO:
Variable   DF        Estimate          Error     Parameter=0

INTERCEP    1        3.422946     0.54805264           6.246
EMOT-SUS    1        0.504836     0.11777245           4.287
JOB-SAT     1        0.316961     0.09222172           3.437
JOB-I       1       -0.227387     0.10026391          -2.266

                                                    Squared
                                  Standardized   Semi-partial
Variable   DF     Prob > |T|         Estimate     Corr Type I

INTERCEP    1         0.0001      0.00000000
EMOT-SUS    1         0.0001      0.32570077      0.15210000
JOB-SAT     1         0.0008      0.27265499      0.04530349
JOB-I       1         0.0248     -0.17604192      0.02805437

                     Squared
                   Semi-partial
Variable   DF  Corr Type II

INTERCEP    1      .
EMOT-SUS    1     0.10022367
JOB-SAT     1     0.06443222
JOB-I       1     0.02805437

                    Variable
Variable   DF       Label

INTERCEP    1   Intercept
EMOT-SUS    1   emotional  sustenance
JOB-SAT     1   job  satisfaction
JOB-I       1   job  involvement

                        The SAS System
   revised model with job involvement as additional predictor

Model: MODEL1
Dependent Variable: LIFE-SAT   life satisfaction

                        Analysis of Variance

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

Model         3      6.82495       2.27498         3.105     0.0402
Error        32     23.44655       0.73270
C Total      35     30.27150

    Root MSE          0.85598     R-square        0.2255
    Dep Mean          5.74000     Adj R-sq        0.1528
    C.V.             14.91257

                     Parameter  Estimates

                     Parameter       Standard     T for HO:
Variable   DF        Estimate          Error     Parameter=0

INTERCEP    1        3.422946     1.15430365           2.965
EMOT-SUS    1        0.504836     0.24809207           2.035
JOB-SAT     1        0.316961     0.19426851           1.632
JOB-I       1       -0.227387     0.21120969          -1.077

                                                    Squared
                                  Standardized   Semi-partial
Variable   DF     Prob > |T|         Estimate     Corr Type I

INTERCEP    1         0.0057      0.00000000              .
EMOT-SUS    1         0.0502      0.32570077      0.15210000
JOB-SAT     1         0.1126      0.27265499      0.04530349
JOB-I       1         0.2897     -0.17604192      0.02605437

                     Squared
                   Semi-partial
Variable   DF  Corr Type II

INTERCEP    1      .
EMOT-SUS    1     0.10022367
JOB-SAT     1     0.06443222
JOB-I       1     0.02605437

                    Variable
Variable   DF       Label

INTERCEP    1   Intercept
EMOT-SUS    1   emotional  sustenance
JOB-SAT     1   job  satisfaction
JOB-I       1   job  involvement
```

The results demonstrate that there is potential for an alternative model to explain life satisfaction and could be verified in a new study. It should also be mentioned that since sample size is input as part of the data step for using correlation matrices, it is possible for a researcher to evaluate the sample size necessary to achieve the desired statistical result and still maintain the necessary values for the adj. R-square, i.e., an indicator of the replicability of the results.

Example 2

The second example of output is drawn from a dissertation assessing the difference in performance on standardized exams completed by Rose (1998) at the University of Arkansas. The study evaluated children born during summer months compared to their colleagues who where born during the academic school year. The results of the study were somewhat inconclusive, even though the theory was solid. Thus, it was our belief that modeling the data using a multivariate method would produce the

results predicted by the theory. The following values were produced by the second program.

MANOVA using SAS IML

*** Multivariate F-test ***

F = 16.366006

*** Univariate F-tests ***

	F_UNIV	PROB
OLSAT	30.757672	< .05
READING	46.437676	< .05
MATH	21.048466	< .05

*** Retrospective and Prospective Power Estimates ***

	POWER	DELTA	POWER-P	DELTA-P
OLSAT	0.72	0.81	0.79	1.28
READING	0.75	0.99	0.60	0.36
MATH	0.69	0.67	0.65	0.54

The results reveal that use of a MANOVA would have produced more positive results and should be considered if a replication study is completed.

EDUCATIONAL IMPACT

The instructional methods in this paper demonstrate several educational benefits for using published data. First, it may encourage researchers and aspiring researchers to use more of the data that is in press to complete preliminary work on new theories. For example, researchers can use components of previous work to test alternative hypotheses to those presented. An additional application could be with the use of large scale data sets where utilizing all the data leads to inflated statistical tests. For example, the Office of Research, Measurement, and Evaluation at the University of Arkansas is using pilot data from the Arkansas educational database to complete pilot studies. The data consists of all standardized test data, both criterion and norm referenced, evaluating K- 12 students annually in Arkansas. However, utilizing all the data leads to issues of non-independence, inflated statistical tests, and results that are hard to interpret with the eclectic elements in Arkansas. Thus, use of pilot data or subsets can be very useful for developing larger studies where additional information will be collected.

Second, power analyses and the computation of effect sizes, a current area of concern in educational research, can be completed for evaluating the appropriateness of sample sizes and possible evidence of the sensitivity of the instruments used in measuring variables. The use of statistical tests has been drawn into question by several editorial boards due to the lack of consideration to meaningfulness or practicality of the statistical results applied to research questions. Reliance on only statistical probability without regard to practicality of the results when applied to the substantive area has led to a move on the part of the American Psychological Association to consider requiring effect sizes.

Third, these methods will also provide a vehicle for further use of meta-analytic studies to compare work from different authors, era's, or fields.

Finally, these methods further demonstrate the impact of understanding the programing mechanisms available in SAS® and their use for instructional purposes.

REFERENCES

Adams, G., King, L., & King, D. (1996). Relationships of Job and Family Involvement, Family Social Support. and Work-Family Conflict with Job and Life Satisfaction. The Journal of *Applied Psychology vol. 8, n4*, 411-420.

Rose, D. (1998). Factors Affecting Success of Summer Children in Elementary School. Unpublished doctoral dissertation, University of Arkansas, Fayetteville.

Sean W. Mulvenon, Ph.D.
241 Graduate Education Building
University of Arkansas
Fayetteville, AR 7270 1
Phone: (501) 575-8727
Fax: (501) 575-2492
E-mail: seanm@comp.uark.edu