REPREG: A REPEATED MEASURES REGRESSION PROGRAM

James T. Love, Boehringer Engelheim
Randy L. Carter, University of Florida

ABSTRACT

The REPREG program is written in the 82.3 release of the Statistical Analysis System (SAS) programming language. REPREG utilizes the macro processing facility (SAS 1982A) and the matrix procedure (SAS 1982B) to perform estimated generalized least squares estimation of linear Random Coefficient Regression (RCR) models. REPREG is useful for estimating an overall regression equation from repeated measures data.

INTRODUCTION

In many regression problems with repeated measures on each of several individuals the usual assumption of independent deviations about an overall regression curve is violated. Thus, ordinary least squares (OLS) estimators are inefficient, even in large samples. The REPREG program performs asymptotically efficient estimation in such cases by computing estimated generalized least square estimates. The vector of generalized least squares (GLS) estimates is a weighted average of OLS estimate vectors from individual regressions. Estimated generalized least squares is the weighted average with estimated weights. The REPREG program computes OLS estimates for each individual, estimates the weighting matrices and computes the weighted average using these estimated weights.

It is assumed that the data arise according to the model

$$\underset{\sim}{y}_i = \underset{\sim}{X}_i \underset{\sim}{\beta}_i + \underset{\sim}{e}_i, \quad i=1, 2, \ldots, n,$$

where $\underset{\sim}{y}_i$ is an $r_i x 1$ vector of observations on the i^{th} of n individual sampling units, $\underset{\sim}{X}_i$ is an $r_i x k$ matrix of observations on k-1 explanatory variables, $\underset{\sim}{\beta}_i$ is a kx1 vector of coefficients unique to the i^{th} individual, and $\underset{\sim}{e}_i$ is an $r_i x 1$ vector of errors. It is further assumed that (1) the $\underset{\sim}{e}_i$ vectors are independently normally distributed with $\underset{\sim}{0}$ mean vector and covariance matrices $\sigma^2 \underset{\sim}{I}_{r_i}$; (2) the $\underset{\sim}{\beta}_i$ vectors are independent drawings from a k-variate normal distribution with mean vector β and nonsingular covarianoe matrix $\underset{\sim}{\Sigma}_{\beta\beta}$; (3) the $\underset{\sim}{\beta}_i$ and $\underset{\sim}{e}_j$ vectors are independent for all i and j; and (4) that several commonly satisfied requirements on n, r_i and $\underset{\sim}{X}_i$ are met. Then, the estimated generalized least square estimator, defined by

$$\hat{\underset{\sim}{\beta}} = \hat{\underset{\sim}{\Omega}}_n (n^{-1} \sum_{i=1}^{n} \hat{\underset{\sim}{W}}_i \underset{\sim}{b}_i),$$

is asymptotically equivalent to the best estimator (i.e., the generalized least squares estimator) as $n \to \infty$ or $\min(r_i) \to \infty$, where $\hat{\underset{\sim}{\Omega}}_n = (n^{-1} \sum_{i=1}^{n} \hat{\underset{\sim}{W}}_i)^{-1}$, $\hat{\underset{\sim}{W}}_i = (\hat{\underset{\sim}{\Sigma}}_{\beta\beta} + \hat{\sigma}^2 (\underset{\sim}{X}_i'\underset{\sim}{X}_i)^{-1})^{-1}$, $\hat{\sigma}^2$ is the pooled MSE from individual regression fits, $\underset{\sim}{b}_i$ is the vector of ordinary least squares estimators for the i^{th} individual, $\hat{\underset{\sim}{\Sigma}}_{\beta\beta} = \underset{\sim}{S}_{bb} - n^{-1} \hat{\sigma}^2 \sum_{i=1}^{n} (\underset{\sim}{X}_i'\underset{\sim}{X}_i)^{-1}$, $\underset{\sim}{S}_{bb} = (n-1)^{-1} \sum_{i=1}^{n} (\underset{\sim}{b}_i - \bar{\underset{\sim}{b}})(\underset{\sim}{b}_i - \bar{\underset{\sim}{b}})'$ and $\bar{\underset{\sim}{b}} = n^{-1} \sum_{i=1}^{n} \underset{\sim}{b}_i$. The user is referred to Swamy (1970, 1971) and Carter and Yang (1982) for more detailed discussions of RCR models and the properties of $\hat{\underset{\sim}{\beta}}$.

The REPREG program computes $\hat{\underset{\sim}{\beta}}$ with $\hat{\underset{\sim}{\Sigma}}_{\beta\beta}$ modified slightly to guarantee positive definiteness (See Carter and Yang, 1982). Several tests of significance that the model parameters are simultaneously or individually zero are performed. Optionally, the user can request similar tests which allow nonzero hypothesized values, tests on linear combinations of model parameters, 95% prediction and confidence intervals about the overall fitted curve and tests of the assumption of equal variance about individual regression curves. Observed, predicted and residual values can be transferred to a new SAS data set, as can the 95% prediction and confidence limits, for use in subsequent SAS procedures such as PROC PLOT.

PROGRAM SPECIFICATIONS

REPREG, like all macros, must be defined in each program in which it is to be used. In an IBM operating system environment this is most easily acccmplished by storing the program listing of REPREG in a permanent card image disk data set and using the following JCL:

```
// EXEC SAS
//SAS.SYSIN DD DSN=data set name
//          DD *
```

After defining REPREG the following statements must be specified each time the program is to be implemented:

```
%LET DATA= data set name;
%LET BYVAR= variable;
```

```
%LET DEPEN= dependent variable;
%LET INDEPEN= independent variable(s);
%LET OPTIONS= list of options;
%LET NULLBETA= hypothesized values;
%LET LINEAR= 'name' coefficients hypothesized
     value;

%REPREG
RUN;
```

THE DATA STATEMENT

```
%LET DATA= data set;
```

'data set' is the name of any SAS data set that would be suitable for use with the GLM procedure.

THE BYVAR STATEMENT

```
%LET BYVAR= variable;
```

'variable' must be a character variable contained in the specified data set. Values of 'variable' must uniquely identify each of the primary sampling units (individuals).

THE OPTIONS STATEMENT

```
%LET OPTIONS= 'optional' 'option2' 'etc.';
```

If no options are desired the semicolon should immediately follow the equal sign.

The following options are available:

PRINT

The estimated generalized least squares estimates produced by REPREG are a weighted average of OLS estimate vectors from individual regressions. These individual regressions are performed within REPREG by the SYSREG[1] procedure.

The PRINT option causes the PROC SYSREG output for each of the individual regressions to be printed. It should be noted that one page of output will be printed for each individual.

VARTEST

Specifying VARTEST results in two tests of the assumption of equal error variance about individual regression curves: (1) the Burr-Foster Q test and (2) Bartlett's test for homogeneity of variance. Bartlett's test is well known and widely used but can be very sensitive to nonnormality of e_i. Furthermore, a zero MSE for any individual excludes a calculation of Bartlett's test. Burr and Foster (1972) derived a test that can be used in the presence of some zero MSE's. Anderson and McClean (1974) discuss the Burr-Foster Q test. Tables of critical values are presented in an appendix to the REPREG documentation. Large values of Q lead to rejection of the homogeneity hypothesis. The tables of critical values are for equal within subject sample sizes, r_i. If the r_i, i=1, 2, ..., n, are not all equal the average degrees of freedom and linear interpolation in the table can be used.

MSEPRINT

MSEPRINT produces a list of MSE's from individual regressions.

LIMITS[2]

By specifying LIMITS a list of observed values, predicted values, residuals, 95% prediction limits (CLI), 95% confidence limits (CLM) and the values of a variable named OUTLIER is printed. OUTLIER is 0 if the observed value is inside and 9999 if the observed value is outside the 95% prediction limits.

OUT[2]

The OUT option results in no printed output but creates a new data set named RESULTS. The variables in RESULTS are named 'dependent variable', ESTIMATE, RESIDUAL, LOW_CLI, UP_CLM, and 'independent variable 1' 'independent variable 2' ... 'independent variable k' and contain observed, predicted, residual, lower prediction limit, upper prediction limit, lower confidence limit, upper confidence limit values, and values of the independent variables, respectively. While the RESULTS data set can be used for a variety of purposes, it is intended primarily for use in PROC PLOT statements to illustrate the fitted overall curve and confidence bands when the model has one independent variable.

THE NULLBETA STATEMENT

The REPREG program will automatically test the null hypothesis that coefficients are individually and simultaneously zero. The NULLBETA statement allows for test of hypotheses that the parameters are specified nonzero values. The general form of this statement is

```
%LET NULLBETA=a1a2...ak/b1b2...
              bk/etc.;
```

where $a_1, a_2, \ldots, a_k$ form one set of hypothesized values for the intercept[3] and coefficients corresponding to k-1 independent variables; $b_1, b_2, \ldots, b_k$ form a second set, etc.

Typically, at most one set of hypothesized values will be desired. If more are requested they must be separated by a /. If the NULLBETA option is requested there must be one hypothesized value for each model parameter. The hypothesized intercept must come first, followed by values for the coefficients on each independent variable in the same order as the independent variables appear in the %LET INDEPEN statement. If no tests of nonzero hypotheses are desired use the statement

```
%LET NULLBETA=;
```

THE LINEAR STATEMENT

The LINEAR subroutine computes estimates and tests of hypotheses for specified linear combinations of the elements of β, the parameter vector for the overall regression curve. Coefficients of and hypothesized values for the linear combinations must be in a statement with the following general form:

%LET LINEAR='NAME1'coefl$_1$coefl$_2$...

coefl$_k$hyp.val$_1$/'Name2'

coef2$_1$coef2$_2$...

coef2$_k$hyp.val$_2$/etc.;

where Name1 is a user assigned name for the first linear combination, coefl...coefl$_k$ are the k coefficients defining the first linear combination, hyp. val. is the hypothesized value of the first linear combination, etc. The names must be enclosed in single quotes and be eight or fewer characters long. There must be exactly one coefficient defined for each model parameter. If several linearly independent linear combinations are defined in the LINEAR statement they will be tested individually and simultaneously. If linearly dependent combinations are specified the simultaneous test will result in an error. Up to k linearly independent combinations may be specified.

EXAMPLE

In a study to relate weight to girth in young calves, weights were obtained after birth and at 4, 6, 7, and 8 weeks. Girths were measured weekly for eight weeks following birth. An initial data set was created by the statements

```
DATA CALVES;
INPUT CALFNO $ TPROT BDATE WT0 WT4 WT6 WT7
     WT8 GR0 GR1-GR8;
CARDS;
```

where CALFNO is the character variable identifying individual calves and WT0, WT4, WT6, WT7, GR0 and GR1-GR8 are measured weights and girths at different times (weeks) after birth. This data set was then modified for use in PROC SYSREG to obtain individual OLS estimates and MSE's. At this point, the observations were transformed. The modification and transformation were accomplished by the statements

```
DATA CALVES; SET CALVES;
LOGWT=LOG(WT0); LOGIRTH=LOG(GR0); OUTPUT;
LOGWT=LOG(WT4); LOGIRTH=LOG(GR4); OUTPUT;
LOGWT=LOG(WT6); LOGIRTH=LOG(GR6); OUTPUT;
LOGWT=LOG(WT7); LOGIRTH=LOG(GR7); OUTPUT;
LOGWT=LOG(WT8); LOGIRTH=LOG(GR8); OUTPUT;
```

Then the following REPREG macro statements were used to estimate an overall relationship between log (weight) and log (girth) from these data:

```
%LET DATA=CALVES;
%LET BYVAR=CALFNO;
%LET DEPEN=LOGWT;
%LET INDEPEN=LOGIRTH;
%LET OPTIONS=VARTEST MSEPRINT LIMITS OUT;
%LET NULLBETA=0 3;
%LET LINEAR='INT' 1 0 0/'SLOPE' 0 1 3;
```

The NULLBETA statement was specified to test the hypothesis that the intercept and slope of the overall log (weight) vs. log (girth) relationship are 0 and 3, respectively. A LINEAR statement for testing these same hypotheses was also specified to illustrate its use for testing linear combinations of parameters.

OUTPUT

Automatic and optional output of REPREG are illustrated below. This output resulted from an RCR analysis of the CALVES data set discussed above. Resulting output is labeled as indicated below.

1. This output resulted from specifying VARTEST in the %LET OPTIONS statement.
2. A portion of the output obtained from the MSEPRINT option.
3. Automatic output which gives the estimated variance-covariance matrix, SB, of the β_i vectors, and n times the estimated variance-covariance matrix of $\hat{\beta}$ (the mean parameter vector estimator) which is labeled OMEGA in the output and $\hat{\Omega}_n$ by Carter and Yang (1982).
4. Automatic output includes sample sizes n, min(r_i), max(r_i) and average r_i, parameter estimates, their estimated standard errors, three test statistics for testing whether the parameters are zero individually, two test statistics for testing whether the parameters are simultaneously zero and approximate p-values for each test. Properties of the estimators and tests are discussed in detail by Carter and Yang (1982) and Swamy (1970, 1971). Briefly, the estimators are consistent as $n\to\infty$, asymptotically unbiased as $\min(r_i)\to\infty$ and asymptotically efficient as $\min(r_i)\to\infty$ or $n\to\infty$. The F tests are appropriate when $\min(r_i)$ is large, the Chi Square tests when n is large, and the t tests when $n\min(r_i)$ is large. A moderately conservative strategy for testing is to always use the F tests. When the X_i matrices are all equal these estimators and tests are those of Rao (1965). In this case the tests are exact and do not require large samples.
5. Optional output obtained from the statement %LET NULLBETA=0 3;. The above comments about the parameter estimates and tests apply to the estimated differences and tests from the NULLBETA statement.
6. Optional output obtained from the statement %LET LINEAR='INT' 1 0 0/'SLOPE' 0 1 3;. The above comments on estimators and tests also apply here.
7. Optional output obtained from the LIMITS option in the OPTIONS statement.

FOOTNOTES

[1] In the 79.6 release of S.A.S., the NOPRINT option of PROC REG does not work properly. Therefore PROC SYSREG is used instead of PROC REG in REPREG.

FOOTNOTES continued

[2]If the number of individuals is large and either the LIMITS option or the OUT option are requested an increased region size will be necessary.

[3]REPREG is not compatible with regression through the origin. An intercept must be included in the model.

REFERENCES

Anderson, V.L. and McClean, R.A. (1974). Design of Experiments: A Realistic Approach. Marcel Decker Inc., New York.

Burr, I.W. and Foster, L.A. (1972). "A Test for Equality of Variance", Department of Statistics, Mimeograph Series No. 282, Purdue University.

Carter, R.L. and Yang, M.C.K. (1982). "Large Sample Inference in Random Coefficient Regression Models". Technical Report No. 168, Department of Statistics, University of Florida.

Rao, C.R. (1965). "The Theory of Least Squares When the Parameters Are Stochastic and Its Application to the Analyses of Growth Curves". Biometrika 52, 447-458.

SAS (1982A). SAS User's Guide: Basics, 1982 edition. SAS Institute Inc., Cary, North Carolina.

SAS (1982B). SAS User's Guide: Statistics, 1982 edition. SAS Institute Inc., Cary, North Carolina.

Swamy, P.A.V.B. (1970). "Efficient Inference in a Random Coefficient Regression Model". Econometrica 38, 311-323.

Swamy, P.A.V.B. (1971). Statistical Inference in Random Coefficient Regression Models. Berlin: Springer-Verlag.

OUTPUT 1: POOLED VARIANCE ESTIMATE

SIGPOOL2
0.00359622

TESTS OF HOMOGENEITY OF VARIANCES

FOR CRITICAL VALUES AND DISCUSSION OF THE BURR-FOSTER Q TEST SEE ANDERSON AND MCLEAN (1974) TABLES ARE GIVEN FOR ALPHA=.01 and .001, FOR 1 TO 20 DEGREES OF FREEDOM, AND FOR 3 TO 60 PRIMARY SAMPLING UNITS

BURR	Q	DF(AVG.)
FOSTER	0.0156995	2.9823
BARTLETT	CHISQ	PROB>CHI
TEST	129.915	0.118498

OUTPUT 2: M.S.E.S FROM INDIVIDUAL REGRESSIONS

ID	M.S.E.
7819	0.00405963
7820	0.00075656
7821	0.00224723
7822	0.00666455
7823	0.00252674
7824	0.000634268
7825	0.00205211
7826	0.00162865
7827	0.000255754
7828	0.00260962
7829	0.000266785
7830	0.00293792
7831	0.00113078
7832	0.00120698
7833	0.000746221
7834	0.0051571

OUTPUT 3: BETA COVARIANCE MATRIX FROM INDIVIDUAL REGRESSIONS

SB	INTER	BETA1
INTER	1.32114	-0.301868
BETA1	-0.301868	0.0690277

N TIMES COVARIANCE MATRIX OF OVERALL BETA ESTIMATE

OMEGA	INTER	BETA1
INTER	4.26979	-0.96583
BETA1	-0.96583	0.218576

OUTPUT 4

STATISTICAL ANALYSIS SYSTEM

NOTE:F IS APPROPRIATE WHEN MIN(RSUBI), THE SMALLEST PRIMARY UNIT SAMPLE SIZE, IS LARGE. CHI IS APPROPRIATE WHEN N, THE NUMBER OF PRIMARY UNITS, IS LARGE. THE DISTRIBUTION OF CHI IS APPROXIMATELY CHISQUARE WITH DF EQUAL TO THE NUMBER OF PARAMETERS IN THE MODEL. T IS APPROPRIATE FOR TESTS OF L'BETA WHEN EITHER N OR MIN(RSUBI) IS LARGE. SEE CARTER AND YANG (1982).

SAMPLE	N	MIN R(I)	MAX R(I)	AVG R(I)
SIZES	113	4	5	4.9823

INDIVIDUAL PARAMETER TESTS
NULL HYPOTHESES:BETA=0

PARM	ESTIMATE	STD ERR	F	PROB>F	CHI	PROB>CHI	T	DF(EST.)	PROB>(T)
INTER	-7.84286	0.194386	1627.87	0.0001	1627.87	0.0001	-40.3469	440.537	0.0001
BETA1	2.82845	0.0439806	4135.95	0.0001	4135.95	0.0001	64.3113	438.681	0.0001

OVERALL PARAMETER TEST
NULL HYPOTHESIS:BETA=ZERO VECTOR
(INTERCEPT NOT TESTED.)

SIMULT	F	PROB>F	CHI	PROB>CHI
TEST	4135.95	0.0001	4135.95	0.0001

OUTPUT 5: TESTS OF BETA=BETA(HYPOTHESIZED)

INDIVIDUAL PARAMETER TESTS
NULL HYPOTHESES:BETA=BETA HYP

PARM	HYPOTH B	EST.DIFF	F	PROB>F	CHI	PROB>CHI	T	DF(EST.)	PROB>(T)
INTER	0	-7.84286	1627.87	0.0001	1627.87	0.0001	-40.3469	440.537	0.0001
BETA1	3	-0.171547	15.2139	0.000164135	15.2139	0.0001	-3.9005	438.681	0.000111041

OVERALL PARAMETER TEST
NULL HYPOTHESIS:BETA=BETA(HYP)

SIMULT	F	PROB>F	CHI	PROB>CHI
TEST	2039892	0.0001	4116540	0.0001

OUTPUT 6: TESTS OF LINEAR COMBINATIONS

NULL HYPOTHESES: L'BETA=HYPOTHESIZED VALUE

COMM	ESTIMATE	HYPOTH	F	PROB>F	CHI	PROB>CHI	T	DF(EST.)	PROB>(T)
INT	-7.84286	0	1627.87	0.0001	1627.87	0.0001	-40.3469	440.537	0.0001
SLOPE	2.82845	3	15.2139	0.000164135	15.2139	0.0001	-3.9005	438.681	0.000111041

SIMULTANEOUS TEST OF LINEAR COMBINATIONS
NULL HYPOTHESIS: VECTOR OF LINEAR COMBINATIONS=VECTOR OF HYPOTHESIZED VALUES

SIMULT	F	PROB>F	CHI	PROB>CHI
TEST	2039892	0.0001	4116540	0.0001

OUTPUT 7

OBSERVED VALUES, PREDICTED VALUES, RESIDUALS,
95% PREDICTION LIMITS, 95% CONFIDENCE LIMITS,
AND OUTLIER INDICATOR

ID	Y	ESTIMATE	RESIDUAL	LOW_CLI	UP_CLI	LOW_CLM	UP_CLM	OUTLIER
7819	4.36945	4.36896	0.000492299	4.2307	4.50721	4.35678	4.38113	0
7819	4.60517	4.60404	0.00112799	4.46838	4.73971	4.5955	4.61258	0
7819	4.85203	4.83711	0.0149248	4.69049	4.98372	4.82702	4.84719	0
7819	4.91265	4.9314	-0.0187411	4.77706	5.08573	4.91945	4.94334	0
7819	5.0626	4.9314	0.131199	4.77706	5.08573	4.91945	4.94334	0
7820	4.33073	4.40642	-0.0756857	4.26951	4.54333	4.39505	4.41779	0
7820	4.62497	4.63854	-0.0135631	4.50206	4.77501	4.63013	4.64695	0
7820	4.78749	4.83711	-0.0496137	4.69049	4.98372	4.82702	4.84719	0
7820	4.95583	4.96214	-0.00631326	4.80494	5.11934	4.9495	4.97478	0
7820	5.07517	5.11104	-0.0358669	4.93795	5.28413	5.09469	5.12739	0
858	4.31749	4.33099	-0.0135011	4.19103	4.47095	4.31794	4.34404	0
858	4.64439	4.65563	-0.0112355	4.51865	4.79261	4.64723	4.66402	0
858	4.77068	4.94681	-0.176125	4.79106	5.10256	4.93452	4.9591	9999
858	4.83628	4.85304	-0.0167585	4.70524	5.00084	4.84267	4.86341	0
858	4.91265	4.99255	-0.0798992	4.83237	5.15274	4.9792	5.0059	0
859	4.47734	4.51592	-0.0385844	4.38094	4.65091	4.50649	4.52535	0
859	4.64439	4.78875	-0.144364	4.64542	4.93209	4.77941	4.7981	9999
859	4.78749	4.80496	-0.0174718	4.66058	4.94934	4.79539	4.81454	0
859	4.8828	4.85304	0.0297615	4.70524	5.00084	4.84267	4.86431	0
859	4.95583	4.9314	0.0244311	4.77706	5.08573	4.91945	4.94334	0
860	4.70048	4.5515	0.148981	4.41648	4.68652	4.54252	4.56048	9999
860	4.82028	4.9314	-0.111114	4.77706	5.08573	4.91945	4.94334	0
860	5.04343	4.94681	0.0966152	4.79106	5.10256	4.93452	4.9591	0
860	5.0876	5.06719	0.0204085	4.8991	5.23528	5.05198	5.08239	0
860	5.15906	5.05242	0.106638	4.88596	5.21888	5.03759	5.06725	0
861	4.52179	4.40642	0.11537	4.26951	4.54333	4.39505	4.41779	0
861	4.66344	4.6895	-0.0260614	4.5513	4.8277	4.68104	4.69796	0
861	4.82028	4.73956	0.0807187	4.59905	4.88007	4.73078	4.74835	0
861	4.86753	4.83711	0.030429	4.69049	4.98372	4.82702	4.84719	0
861	4.96981	4.94681	0.0230034	4.79106	5.10256	4.93452	4.9591	0
862	4.58497	4.51592	0.0690462	4.38094	4.65091	4.50649	4.52535	0