**Paper # ST-12**

# A Macro for Computing a Goodness of Fit Statistic for Linear Mixed Models

**Jean G. Orelien, Analytical Sciences Inc., Durham, NC**

**Abstract:**

In the SAS System, PROC MIXED offers users the possibility to perform analysis of complex data where assumptions of traditional analysis of variance (ANOVA) methods such as homogeneity of variance or independence of error terms might be violated. Thus, this procedure can be used to analyze data where the observations are assumed to come from a normal distribution but are correlated such as in longitudinal studies or studies where the data are collected from clusters (center, school, city). Unfortunately, in the linear mixed models, there are few tools available for checking adequacy of the model. In this paper, we present a macro to compute a goodness-of-fit statistic denoted model concordance correlation that was proposed by Vonesh et al. (1996). One advantage of this statistic is that it is similar to the $R^2$ used in traditional ANOVA and can be used to assess the adequacy of the assumed mean and covariance structure. PROC IML is used to perform the computations. This paper should be accessible to anyone who is familiar with linear regression methods.

## I Introduction

In linear mixed models, few diagnostic tools are available for assessing adequacy of the model. Two statistics that are often used and available in the SAS System are the Akaike's information criterion (AIC) and the Bayesian information criterion. These statistics can be useful when comparing the fit of several models to the same data. However, the problem with these two statistics for assessing goodness of fit is that they may not be intuitively interpretable in that they do not have well defined endpoints corresponding to a perfect fit or a complete lack of fit. That is given a linear mixed model, these two statistics do not provide us a sense of how good this model (on some scale) is and how much improvement might be needed. For example, in traditional ANOVA an $R^2$ of 0.8 can be interpreted as "80% of the variation in the dependent variable can be explained by the independent variable(s)".

In this paper, we present a goodness-of-fit statistic denoted the concordance correlation coefficient (CCC) that is comparable to traditional $R^2$ in linear regression. This goodness-of-fit statistic was first introduced by Lin (1989) to assess the measure to which 2 sets of values agree. Later Vonesh et al. (1996) proposed to use CCC to compare the degree to which observed and expected values agree for a large class of models that include the mixed-effect model.

In section 2, we present the linear mixed effect model and its assumption. We give the formula and interpretation of CCC in section 3. A macro to compute CCC is discussed in section 4. An example showing the use of this macro is given in section 4. Advantages and limitations of CCC are discussed in section 5. The performance of CCC is discussed in section 6.

## II. Model Assumptions

Assume that we have the following model:

$$\boldsymbol{Y_i = X_i\beta + Z_i d_i + e_i}$$

Where:
$\boldsymbol{Y_i}$ is an $\boldsymbol{n_i x 1}$ vector of observations from the ith cluster. For a longitudinal study, this would correspond to the vector of observations from the ith individual.

$\boldsymbol{X_i}$ denotes an $\boldsymbol{n_i xp}$ fixed effects design matrix for the ith subject
$\boldsymbol{\beta}$ is a px1 vector of unknown, constant, fixed effect parameter estimates
$\boldsymbol{Z_i}$ denotes an $\boldsymbol{n_i xq}$ random effects design matrix for the ith subject
$\boldsymbol{d_i}$ is a $\boldsymbol{qx1}$ vector of unobservable random effects or subject effect coefficients
$\boldsymbol{e_i}$ denotes an $\boldsymbol{n_i x1}$ vector of unobservable within-subject error terms.

It is also assumed that $\boldsymbol{d_i}$ has a multivariate normal distribution $N_q(\mathbf{0},\ \boldsymbol{D})$ independent of $\boldsymbol{e_i}$ which has a multivariate distribution $N_{n_i}(\mathbf{0},\ \sigma^2\boldsymbol{V_i})$.
Thus, we have:

$$E\begin{bmatrix}\boldsymbol{d_i}\\ \boldsymbol{e_i}\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\ V\begin{bmatrix}\boldsymbol{d_i}\\ \boldsymbol{e_i}\end{bmatrix}=\begin{bmatrix}\boldsymbol{D} & \mathbf{0}\\ \mathbf{0} & \sigma^2\boldsymbol{V_i}\end{bmatrix}$$

and for $\boldsymbol{i' \neq i}$, we also have:
$\boldsymbol{Cov[d_i, d_{i'}] = 0,\ Cov[e_i, d_{i'}] = 0,\ Cov[e_i, e_{i'}] = 0}$
Where:
$\boldsymbol{D = Var(d_i)}$ is the qxq covariance matrix of the random effects
$\sigma^2$ is an unknown scalar within-subject error variance parameter
$\boldsymbol{Var(e_i) = \sigma^2 V_i}$ is the covariance matrix of the random deviations about the ith subject's random regression line.

## III The concordance correlation coefficient

The CCC is given by the formula:

$$CCC = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)'(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y}1_i)'(y_i - \bar{y}1_i) + \sum_{i=1}^{n}(\hat{y}_i - \hat{y}1_i)'(\hat{y}_i - \hat{y}1_i) + N(\bar{y} - \hat{y})^2}$$

Where:
*n* is the number of clusters (or subjects )
$y_i$ is the vector of observed values for the ith cluster
$\hat{y}_i$ is the vector of predicted values for the ith cluster
$\bar{y}$ is the average of the observed values in the ith cluster
$\hat{y}$ is the average of the predicted values in the ith cluster
*N* is the total number of observations
The CCC can be viewed as a statistic measuring the degree of agreement between the predicted and observed values. If there is perfect agreement between observed and predicted, one would expect a scatter plot of observed versus predicted to be equivalent to the identity line. Recall in contrast that the interpretation of $R^2$ is the percentage of variation in the dependent variable that can be explained by the independent variables.

## IV A macro to compute the CCC

A macro to compute the CCC is provided in the appendix. This macro assumes that PROC MIXED is used and that the model meets the following requirements:
1) It is assumed that there is a random statement.
2) There is an intercept term for the fixed effects and random effects

Two inputs for the program are data sets that are generated by PROC MIXED These data sets are the data set for the random effect parameters and the data set for the fixed effect parameters. One can easily create these data sets using ODS. Example:

```
ods output solutionf=fixed;
ods output solutionr=random;
```

The macro requires input for the following macro variables:

```
DEP:  Specify the dependent variable

FIXED_EFFECTS:

This macro variable contain the
variables for the fixed effects

RANDOM_EFFECTS:
The macro variable random effects
contain the variables for the random
effects (excluding the intercept)

NUM_RANVAR: Numbers of random
            variables(excluding the
            intercept)

INDATA:     the macro variable INDATA
            identifies the SAS data set
            used in the analysis

IDVAR:      ID variable;

PARMF:      Identifies the data set
            that contains the fixed
            effect parameters

PARMR:      Identifies the data set
            that contains the random
            effect parameters;
```

## V. Advantages and Disadvantages of the CCC

The CCC offers the advantage that it's values lie between -1 and 1 with a perfect fit corresponding to a value of 1 and a lack of fit corresponding to values less than or equal to 0. Thus, it has an intuitive interpretation. Furthermore, in the case of linear regression with an intercept, the CCC is directly related to $R^2$ with the relation $CCC = 2R^2/(R^2+1)$. Unlike the AIC or the BIC, the CCC does not compare the model at hand to other models, thus it does not require that other models be fitted.

CCC as we define it will answer the question of whether the model specified (including the fixed effect and the covariance structure) fits the data. One may be interested in asking separately the questions: 1) Is the response function adequate (i.e., fixed effects)? and 2) Is the covariance structure for the random effects adequate? The adequacy of the covariance structure specified by the user in PROC MIXED cannot be directly tested by CCC. However, a test described in Vonesh et al. can easily be implemented in the SAS System.

It should also be noted that CCC does not adjust for the number of parameters in the model. Similarly to the traditional $R^2$, CCC will increase with increasing number of parameters in the model. But one can performed the same type of adjustment for the number of parameters in the model that is done with $R^2$.

One of the major disadvantage of CCC is it's unstated assumptions. One of the underlying assumptions are that any two predicted values or any two observed values are independent. Clearly in the case of longitudinal studies, this is not true. Since observations from two subjects are correlated. Another assumption of CCC is that $(y, \hat{y})$

come from a bivariate normal distribution. However, Lin(1989) showed that CCC was robust to departure from the assumption of bivariate normal.

## VI Performance of CCC

We conducted a simulation based on data reported by Potthoff and Roy (1964) in order to assess the performance of the CCC. In the data from Potthoff and Roy, the outcome of interest is the distance in millimeters between the pituary and the pterygomaxillary fissure. Repeated measures were taken at ages 8, 10, 12 and 14 on 17 boys and 16 girls. We simulated data in which the outcome was in terms of fixed effects a function of age, gender and age by gender interaction. Age was the only random effect term used in the simulated data. We generated 2000 runs of the data for sample sizes of 10, 20 and 30 subjects.

Table I gives the values of CCC and AIC from the full model. In table II, there is little change in the values of CCC and AIC from the reduced model that doesn't include the age by gender interaction term. In Table III, removing gender and age by gender interaction, we obtained average CCC that were similar to the full model suggesting that the reduced model with these terms removed is not worst than the full model. However, based on the differences between the averages of the AIC for this reduced model and the full model, we would conclude that the full model gives a better fit. In a model with only fixed effect intercept and random intercept as covariates, the value of CCC obtained was closed to 0.5 not 0.

Since high values of CCC were obtained even when important terms were missing from the model, one needs to be careful in interpreting large values of CCC. On the other hand, a small value of CCC may indicate some problems in the adequacy of fit of the model.

**Table 1. Results from the full model (true model)**

| No. of Subjects | Average CCC | Average AIC |
|---|---|---|
| 10 | 0.88 | 160.83 |
| 20 | 0.89 | 322.65 |
| 30 | 0.89 | 485.01 |

**Table 2. Results from the reduced model (no age by gender interaction)**

| No. of Subjects | Average CCC | Average AIC | AIC full - AIC Reduced |
|---|---|---|---|
| 10 | 0.88 | 161.88 | -1.47 |
| 20 | 0.89 | 324.43 | -1.78 |
| 30 | 0.89 | 487.40 | -2.02 |

**Table 3. Results from the second reduced model (no gender or age by gender interaction)**

| No. of Subjects | Average CCC | Average AIC | AIC full - AIC Reduced |
|---|---|---|---|
| 10 | 0.90 | 171.72 | -10.93 |
| 20 | 0.91 | 341.20 | -18.55 |
| 30 | 0.91 | 511.28 | -25.85 |

## VII Conclusion

Few diagnostic tools are available for checking the adequacy of the model fit in PROC MIXED. We have provided a macro to compute a statistic that can be compared to the $R^2$ of traditional ANOVA. Based on the results of our simulation we suggest using this statistic in conjunction with other goodness of fit statistics. A high value of this statistic (close to 1) may indicate a good fit. But it is possible that other important covariates might be missing from the model even with large values. A value of this statistic close to 0.5 seem to be an indication of a poor fit.

**Reference:**


Lin, LI (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics **45**, 255-268.

Vonesh EF, Chinchili VM, Pu K. (1996). Goodness-of-fit in generalized nonlinear mixed-effect models. Biometrics **52**: 572-587.

Potthoff RF and Roy SN (1964). A generalized multivariate analysis of variance model useful especially for growth curve problem. Biometrika **51**, 313-326.


**Appendix**

```
/******
This macro computes the model
```

```
concordance correlation coefficient
suggested by Vonesh et al. (1996) for a
mixed model

The macro has the following
assumptions:
1) It is assumed that there is a random
statement.
2) There is an intercept term for the
fixed effects and random effects

Among other things the macro will also
create a data set called PREDICTED that
contains the predicted value

**********/

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

The following macro variables need to
be specified:

%let dep=dbp;   Specify the dependent
variable

%let fixed_effects=diet time timesq
timecu;
 the macro variable fixed effects
contain the variables for the fixed
effects
%let random_effects=time timesq;
 the macro variable random effects
contain the variables for the random
effects

%let num_ranvar=2;

 Numbers of random variables

%let indata=diet;

 the macro variable indata identifies
the SAS data set used in the analysis

%let idvar=shortid;

 ID variable

%let parmf=fixed;

 parmf is the data set that contains
the fixed effect parameter

%let parmr=random;

 parmr is the data set that contains
the random effect parameters

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

%macro pred;

proc sort data=&parmr; by &idvar;

data &parmr;
set &parmr;
by &idvar;
n=int(_n_/%eval(&num_ranvar+1));
call symput('n_units', n);

data temp;
set &parmr;
by &idvar;
if first.&idvar;
keep &idvar n;

proc sort data=&indata; by &idvar;

data &indata;
merge &indata (in=a) temp(in=b)
;
by &idvar;
dummy=1;
if a and b;

%put &n_units;

proc means data=&indata noprint;
var dummy;
by &idvar;
output out=size(drop=_type_ _freq_)
sum=;

proc iml;
use size;
read all var{dummy} into clustersize;
use &parmf;
read all var {estimate} into beta;
use &parmr;
read all var {estimate} into d;
use &indata;
read all var  {dummy &fixed_effects}
into x;
read all var {dummy &random_effects}
into z;
read all var {&dep} into y;

do i=1 to &n_units;
iplus=i + 1;
iminus=i - 1;
start_di=(%eval(&num_ranvar+1))*(i-
1)+1;
end_di=(%eval(&num_ranvar + 1))*i;
if i=1 then do;
start_xi=1;
end_xi=clustersize[1, 1];
end;
if i > 1 then do;
clustersizei=clustersize[1:iminus, ];
start_xi=clustersizei[+, +]+1;
end_xi=(start_xi-1)+clustersize[i, 1];
end;

di=d[start_di:end_di, ];
xi=x[start_xi:end_xi, ];
zi=z[start_xi:end_xi, ];
pi=xi*beta + zi*di;
ni=j(nrow(pi),1, i);
predi=ni||pi;
p=p//pi;

pred=pred//predi;
```

```
end;

varname={'n' 'pred'};

/* compute grand mean for dependent
value; */

ybar=y[:];
yhat=p[:];

ncol_yhat=nrow(p);
ncol_y=nrow(y);

/* now compute rc; */

do i=1 to &n_units;

iplus=i + 1;
iminus=i - 1;

if i=1 then do;
start_xi=1;
end_xi=clustersize[1, 1];
end;
if i > 1 then do;
clustersizei=clustersize[1:iminus, ];
start_xi=clustersizei[+, +]+1;
end_xi=(start_xi-1)+clustersize[i, 1];
end;
yi=y[start_xi:end_xi]; * could change
this into start_yi ;
yhati=p[start_xi:end_xi ];
ni=j(nrow(yi),1, i);
ssqyi_yhati=( t( (yi - yhati) ) )*(yi -
yhati);
ssqyi_ybar=(t( (yi -  (j(nrow(yi), 1,
ybar)) ) ))*( yi - (j( nrow(yi), 1,
ybar)) );
ssqyhati_yhat=(t( (yhati -j(nrow(yi),
1, yhat)) ))*(yhati -(j(nrow(yi), 1,
yhat)));

ssq1=ssq1//ssqyi_yhati;
ssq2=ssq2//ssqyi_ybar;
ssq3=ssq3//ssqyhati_yhat;

end;
num=ssq1[+,+];
den1=ssq2[+,+];
den2=ssq3[+, +];
den3=(clustersize[+,+])*((ybar -
yhat)**2);

rc=1 - ( (num)/(den1+den2+den3) );

print "Model Conconcordance
Correlation" rc;

print "Grand Mean " ybar ;

print "Means of all predicted" yhat;

create predicted from pred[colname={n
pred}] ;
append from pred;

quit;

%mend pred;

%pred; run;
```