Paper 2424-2018

Estimation of correlation coefficient in data with repeated measures

Katherine Irimata, Arizona State University; Paul Wakim, National Institutes of Health; Xiaobai Li, National Institutes of Health

ABSTRACT

Repeated measurements are commonly collected in research settings. While the correlation coefficient is often used to characterize the relationship between two continuous variables, it can produce unreliable estimates in the repeated measure setting. Alternative correlation measures have been proposed, but a comprehensive evaluation of the estimators and confidence intervals is not available. We provide a comparison of correlation estimators for two continuous variables in repeated measures data. We consider five methods using SAS/STAT® software procedures, including a naïve Pearson correlation coefficient (PROC CORR), correlation of subject means (PROC CORR), partial correlation adjusting for patient ID (PROC GLM), partial correlation coefficient (PROC MIXED), and a mixed model (PROC MIXED) approach. Confidence intervals were calculated using the normal approximation, cluster bootstrap, and multistage bootstrap. The performance of the five correlation methods and confidence intervals were compared through the analysis of pharmacokinetics data collected on 18 subjects, measured over a total of 76 visits. Although the naïve estimate does not account for subject-level variability, the method produced a point estimate similar to the mixed model approach under the conditions of this example (complete data). The mixed model approach and corresponding confidence interval was the most appropriate measure of correlation as the method fully specifies the correlation structure.

INTRODUCTION

The correlation coefficient ρ is often used to characterize the linear relationship between two continuous variables. Estimates of the correlation (r) that are close to 0 indicate little to no association between the two variables, whereas values close to 1 or -1 indicate a strong association. The sign of the correlation estimate, either positive or negative, reflects the direction of the relationship. In repeated measures data, in which multiple measurements are collected on the same subject, one must exercise caution when calculating the correlation. This type of data has variation between the subjects as well as variation within the repeated measurements on each subject which needs to be accounted for.

In this article, we provide a comparison of correlation measures for two continuous variables and demonstrate the performance of methods on pharmacokinetics data for 18 subjects (Kim et al., 2017). Each patient in the study has between two and eight visits, with measurements collected at each visit, for a total of 76 visits. Information collected on each patient may include the area under the curve from time 0 to 4 hours (AUC04), area under the curve from time 0 to infinity hours (AUCInf), halflife (hours), volume of the distribution (mL), weight (kg), body surface area (BSA, m^2), clearance (mL/hour), and the peak plasma concentration of the drug (Cmax, nM). Figure 1 displays a scatterplot of the variables. We are primarily interested in understanding how various pharmacokinetics covariates are associated with AUC04 (area under the curve from time 0 to 4 hours) and AUCInf (area under the curve from time 0 to infinity hours).

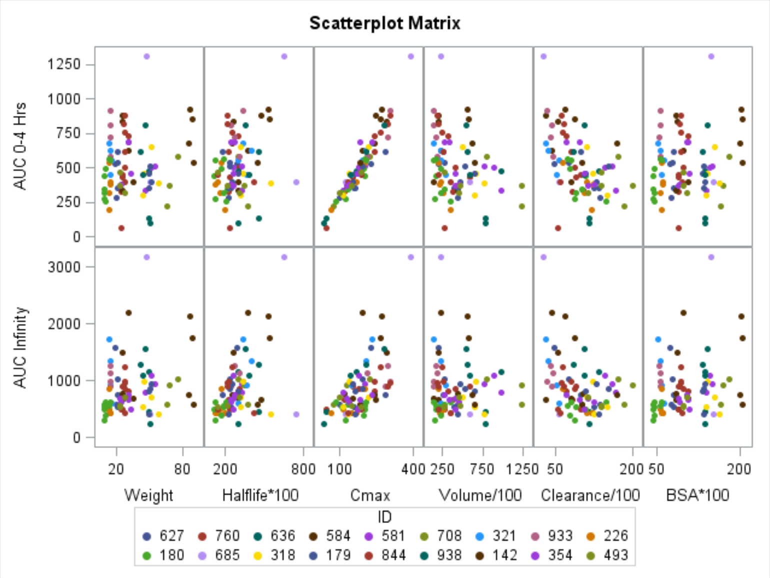


Figure 1. Scatterplot of AUC04 and AUCInf vs pharmacokinetics measurements

Each correlation measure and corresponding confidence interval are introduced, as well as the procedure to calculate the correlation measure in SAS. The objective of our study is to identify a measure that is best for describing correlation in repeated measures data. We discuss the appropriateness of each measure and provide recommendations. We conclude with a further discussion of inference methods for the selected correlation coefficient.

ESTIMATION METHODS IN SAS

Five methods for calculating the point estimate of correlation were evaluated and compared. Two subject level summary approaches (the naïve approach and correlation of subject means) were investigated, as well as three modeling approaches (partial correlation adjusting for subject effect, partial correlation coefficient, and mixed model).

NAÏVE APPROACH

Although repeated measures data are complex due to the differences that exist between subjects, one approach to evaluate the correlation is to assume the observations are independent. While this ignores the inherent grouping structure, we consider this method as a possible approach (Bland and Altman, 1994). SAS PROC CORR can be used to evaluate the Pearson correlation between two specified variables. The syntax is shown below:

```
proc corr data=dta;
      var x y;
run;
```

CORRELATION OF SUBJECT MEANS

The naïve approach ignores the correlation between the repeated measurements on the same subject. To account for the variation between subjects, we consider a second approach that summarizes the correlation at the subject level (CSM). We can estimate the correlation using the Pearson correlation coefficient on the subject averages ($\bar{X}$ and $\bar{Y}$) between two measurements. It is important to note that this greatly reduces the effective sample size to the number of subjects in the dataset. For example, in the data we used, it reduced the number of observations from 76 to 18. PROC SUMMARY can be used together with PROC CORR to calculate the Pearson correlation of the subject means:

```
proc summary data=dta;
      var x y;
      by ID;
      output out=avg mean(x)=avg_x mean(y)=avg_y;
run;
proc corr data=avg;
      var avg_x avg_y;
run;
```

PARTIAL CORRELATION ADJUSTING FOR PATIENT EFFECT

The third proposed method evaluates the partial correlation between two variables after adjusting for the subject (PCA). We can partial out the subject effect using regression, and then calculate the Pearson correlation on the residuals (Christensen, 2011). SAS PROC GLM can be used to calculate the partial correlation on continuous and categorical variables for a third covariate Z (such as ID) as:

```
proc glm data=dta;
      class z;
      model x y = z;
      manova/printe;
quit;
```

The SAS option MANOVA/PRINTE displays the error sums of squares cross product (SSCP) matrix and the partial correlation matrix. As an alternative, SAS PROC CORR also has a partial statement that can be used, although this is limited to use for continuous variables (in the case the investigator wants to partial out a variable other than subject effect, such as height, weight, etc.):

```
proc corr data=dta;
      var x y;
      partial z;
run;
```

PARTIAL CORRELATION COEFFICIENT

The fourth method evaluated is the partial correlation coefficient (PCC) proposed by Lipsitz (2001). This measure quantifies the impact of one variable on a second variable. This is not a proper measure of correlation as it does not have the property $\rho_{XY} = \rho_{YX}$. However, it can be used to describe the relative strength of the association between the two variables. An appropriate repeated measures model of the response in terms of the predictor(s) is fit and the correlation is calculated as

$$\hat{\rho}_{\text{PCC}} = \frac{Z/\sqrt{m}}{\sqrt{(1+Z^2/m)}} \text{ where } Z = \hat{\beta}\Big/\sqrt{\widehat{var}(\hat{\beta})}$$

for m subjects, where Z is the maximum likelihood Wald statistic. To calculate the partial correlation coefficient in SAS, we first fit the repeated measures model. This can be performed using PROC MIXED, which can introduce a random effect for each subject. The syntax is shown below:

```
proc mixed data=pk method=ml;
      class ID visit;
      model x = y / solution;
      repeated visit /subject=ID type=CS;
      ods output SolutionF=SolutionF;
run;
```

In PROC MIXED, the CLASS statement identifies categorical variables. This includes the patient ID and the visit number. In the model statement, the selected response (X) is specified as a function of the selected predictor (Y). The REPEATED statement controls the covariance matrix at the subject level (SUBJECT=ID) for the VISIT using a compound symmetry correlation structure (TYPE=CS). PROC MIXED produces output with the model fit and model estimates. The ods output table SOLUTIONF contains the model estimates of the fixed effects $\hat{\beta}$ and the variances $\widehat{var}(\hat{\beta})$. PROC IML can be used to read in the model estimates and variances from SOLUTIONF and calculate the partial correlation coefficient estimate as described above.

MIXED MODEL

The final method investigated is a mixed model approach (MM) proposed by Hamlett, Ryan, and Wolfinger (2004). This method calculates the correlation estimate from the within subject variance matrix $V_i = var(Y_i)$ for $i = 1, \dots, n$. For this method, we assume that

$$\begin{bmatrix} X_{ij} \\ Y_{ij} \end{bmatrix} \sim BVN\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \Sigma\right),$$

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_X \sigma_Y \rho_{XY} \\ \sigma_X \sigma_Y \rho_{XY} & \sigma_Y^2 \end{bmatrix}.$$

We fit a mixed model with an unstructured correlation structure on the random effects γ_i and the residuals ϵ_i. The form of the mixed model is

$$(\text{Response})_{\text{i}} = X_i \begin{pmatrix} \mu_X \\ \mu_X - \mu_Y \\ 0 \end{pmatrix} + Z_i \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} + \epsilon_i$$

with random effects $(\gamma_1, \gamma_2)' \sim N((0,0)', G)$ and $\epsilon_i \sim N(\mathbf{0}, R)$ where

$$G = \begin{pmatrix} \sigma_X^2 \rho_X & \sigma_{XY} \delta \\ \sigma_{XY} \delta & \sigma_Y^2 \rho_Y \end{pmatrix}, R = I_{m_i} \otimes \begin{pmatrix} \sigma_X^2 (1 - \rho_X) & \sigma_{XY} (1 - \delta) \\ \sigma_{XY} (1 - \delta) & \sigma_Y^2 (1 - \rho_Y) \end{pmatrix}.$$

The design matrix X_i is dimension $2m_i \times 3$ and the random effects matrix Z_i is dimension $2m_i \times 2$. In Hamlett et al.'s method (2004), the correlation structure is clearly defined by the mixed model set up (Figure 2).

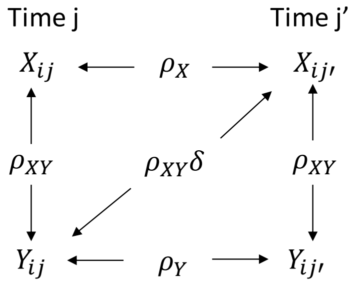


Figure 2. Correlation structure for the mixed model approach

To calculate the correlation using the mixed model approach, the data must first be converted from wide (multivariate) to long (univariate) format. The SAS code below converts the data with two variables (X and Y) into one variable (Response). The variable Vtype denotes which variable value is contained in the line (1 = X, 2 = Y). SAS PROC MIXED can then be used to fit the repeated measures model with the new variables Response and Vtype:

```
data data_long (drop=x y i);
      set data (keep=ID Visit x y);
      array var[2] x y;
      do i = 1 to 2;
            Vtype = i;
            Response = var[i];
            output;
      end;
run;
proc mixed data=data_long method=ml;
      class ID vtype visit;
      model response = vtype / solution ddfm=kr;
      random vtype / type=un subject=ID v vcorr;
```

```
        repeated vtype / type=un subject=visit(ID);
        ods output VCorr=VCorr ConvergenceStatus=CS;
    run;
```

The RANDOM statement specifies the random effect terms that will be included in the mixed model, and TYPE= defines the type of covariance matrix that relates the random effect terms. In this case, two random effect terms (γ_1 and γ_2) are defined for the two levels of the Vtype variable with an unstructured covariance structure. The SUBJECT= option defines the random effects for each subject (specified using the ID variable). The REPEATED statement sets up the covariance structure for the residuals, which is also specified to be unstructured through the option TYPE=UN. The ODS OUTPUT statement saves two datasets. The dataset VCorr contains the correlation matrix between visits for each subject and CS saves the convergence status to check if the mixed model converged. Convergence issues may occur due to variables of different magnitudes. Rescaling may resolve some convergence issues.

CONFIDENCE INTERVALS

To compare the correlation estimates, we calculate various confidence intervals as a method of inference. Three confidence intervals were considered including a normal approximation using Fisher's z transformation and two bootstrap methods. The corresponding 95% confidence interval was calculated for each estimation method as appropriate.

NORMAL APPROXIMATION

A normal approximation can be used to calculate the confidence interval. The correlation coefficient can be transformed using a Fisher's z transformation

$$z = 0.5\log\left(\frac{1+r}{1-r}\right)$$

with approximate variance $V(z) = 1/(n-3)$ (Shen and Lu, 2006). As the sample size increases, the distribution of z approaches normality. The confidence interval can be calculated indirectly by finding the confidence interval of ξ and then using a transformation to obtain the confidence interval of the correlation coefficient ρ:

$$(\xi_L, \xi_U) = \left(z_r - z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{1}{n-3}}, z_r + z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{1}{n-3}}\right)$$

$$(r_L, r_U) = (\tanh(\xi_L), \tanh(\xi_U)) = \left(\frac{\exp(2\xi_L)-1}{\exp(2\xi_L)+1}, \frac{\exp(2\xi_U)-1}{\exp(2\xi_U)+1}\right)$$

The normal approximation confidence interval was used to compare correlation estimates for the naïve, correlation of subject means (CSM), and the partial correlation coefficient (PCA and PCC) approaches.

BOOTSTRAP APPROACHES

Bootstrapping uses random sampling with replacement to estimate the sampling distribution of a test statistic and can be used to form confidence intervals. In general, bootstrap samples are generated by drawing samples from Y with replacement and calculating the value of interest on the sample ($\hat{\rho}^*$). For B bootstrap samples, the bootstrap estimate is calculated as

$$\hat{\rho}^* = \frac{\sum_{i=1}^{B}\hat{\rho}^{*(i)}}{B}.$$

Bootstrapping produces biased estimates. The bias of a bootstrap statistic can be determined by $Bias(\hat{\rho}^*) = \hat{\rho}^* - \hat{\rho}$. Confidence intervals for bootstrap estimates can be formed using various approaches, including percentile, bootstrap t, bias corrected, and bias corrected and accelerated. We focus on the bias corrected confidence interval. Cluster and multistage bootstrap approaches, which vary based on the resampling method, were considered to evaluate the correlation.

Cluster Bootstrap

The cluster bootstrap approach samples from the highest-level clusters with replacement. This method ignores any variation at the lower level (within a cluster). In the case of the pharmacokinetics data, individuals were successively sampled with replacement and the correlation coefficient was estimated for each sample using the various methods described above. The lower level, the patient visits over time, were not resampled. In our study, cluster bootstrapping was used to compare correlation estimates for the correlation of subject means (CSM), partial correlation adjusting for patient effect (PCA), partial correlation coefficient (PCC), and the mixed model (MM) approaches.

Multistage Bootstrap

For multistage bootstrapping, resampling is performed for each level of the hierarchy. The highest-level clusters are sampled with replacement first. Within each cluster, the observations are then selected with replacement. In the pharmacokinetics data, each bootstrap sample contains individuals selected with replacement as well as a resampled set of visits for each patient. Multistage bootstrapping was used to compare correlation estimates for the correlation of subject means (CSM), partial correlation adjusting for patient effect (PCA), and the mixed model approaches (MM).

RESULTS

Six pairs of pharmacokinetics variables were selected and evaluated using the various correlation coefficient estimation methods and the corresponding confidence intervals (AUC04 with halflife, volume, and weight; AUCInf with BSA, clearance, and Cmax). The estimates and confidence intervals are displayed in Figure 3.

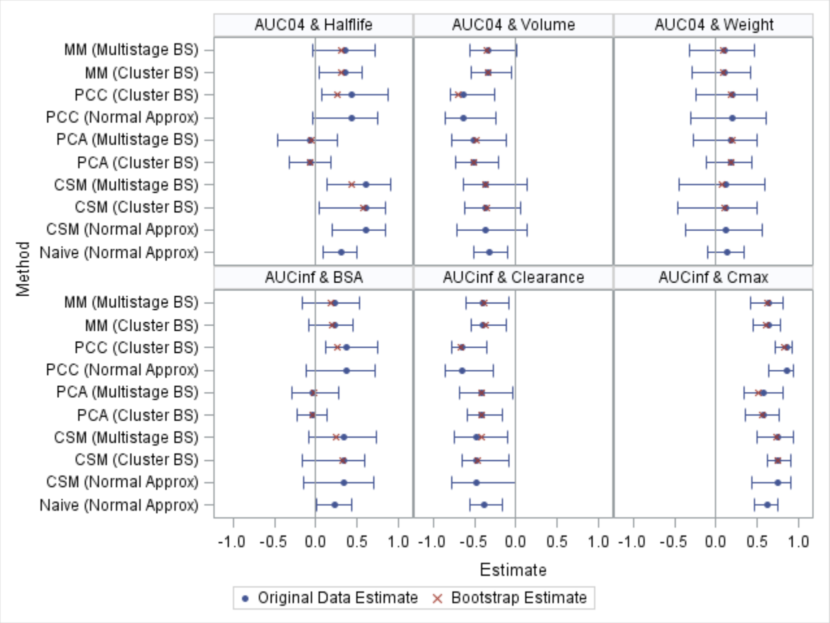


Figure 3. Correlation estimates and confidence intervals

From the plots in Figure 3, we see that the naïve and mixed model approaches tended to produce similar correlation estimates for the pharmacokinetics data. In addition, the naïve and MM methods tended to produce weaker correlation estimates (closer to 0) while the PCC estimates were relatively stronger (closer to 1 or -1). In some cases (such as with AUC04 & Halflife and AUCInf & BSA), PCA estimates varied substantially from other correlation estimates. The naïve approach tended to have the narrowest confidence interval while CSM (normal approximation) tended to have the widest confidence interval. Multistage bootstrapping produced wider confidence intervals compared to cluster bootstrapping. CSA (multistage bootstrap) and PCC (cluster) tended to have the largest bootstrap bias. In four of the cases (AUC04 & Halflife, AUC04 & Volume, AUCInf & BSA, AUCInf & Clearance), the confidence intervals suggested conflicting results in terms of statistical significance. Although not reported in the figure, some mixed models had convergence issues due to scale or the complexity of the model (convergence rates ranged from 96.38% to 99.99%).

DISCUSSION AND RECOMMENDATIONS

We review the correlation methods investigated to determine the most appropriate correlation estimate for repeated measures data. The correlation of subject means approach only considers correlation between the means, and is not able to identify the correlation between the covariates at each visit. This limits the CSM measure from accounting for within subject variation. The PCA approach adjusts for the subject effect (using the variable ID) in each covariate separately, which many not maintain the overall structure of the data. The partial correlation coefficient is not truly a correlation measurement as it measures the predictive strength of one covariate on the other. Switching the two covariates in the repeated measures model results in completely different PCC estimates, and as seen in Figure 3, in some cases the PCC estimates tended to vary greatly from other correlation estimates. Overall, the naïve and mixed model approaches produced similar correlation estimates for this data, although in general the naïve approach does not account for subject level variability. Under the conditions of the pharmacokinetics example, it appears that the naïve approach provides a reasonable point estimate of correlation, although the confidence interval should not be used.

Based on the review of the selected correlation methods and the results shown in Figure 3, we recommend the mixed model approach to obtain the correlation estimate and corresponding confidence interval. While this method should be further investigated using simulation, the mixed model estimation method appears to be the most appropriate measure as it fully specifies the correlation structure. Model assumptions, including normality, should be checked when fitting the mixed model. The bootstrap method for the confidence interval should be selected based on the data structure. For example, in data where the measurements are independent given the subject, multistage bootstrap is preferred. Cluster bootstrap should be used when the measurements within a subject have a particular correlation structure. A p-value for the correlation is not provided with the mixed model approach. Bootstrap p-values should be calculated with caution, as they need to be generated under the null hypothesis which should be reflected in the bootstrap sampling scheme.

ADDITIONAL DISCUSSION OF THE MIXED MODEL APPROACH

From the comparison of various correlation coefficients for repeated measures data, we identify the mixed model approach (Hamlett et al. 2004) as the most appropriate measure. We provide further details on this recommended approach by introducing methods of inference. This includes a normal approximation, which uses the delta method, as well as a further discussion of the cluster bootstrap for inference using both a confidence interval and p-value.

We select two pairs of pharmacokinetics variables, AUC04 and Clearance (scaled down by a factor of 100 to improve the convergence of the mixed model approach) and log(AUC04) and BSA, to evaluate the correlation. Prior to implementing the mixed model approach, we verify that the model assumptions are met. The estimated correlations for AUC04 and Clearance/100 and log(AUC04) and BSA are -0.5341 and 0.0519, respectively. Thus, we evaluate a relatively strong correlation and a relatively weak correlation in repeated measures data.

NORMAL APPROXIMATION (CONFIDENCE INTERVAL AND P-VALUE)

The confidence interval and p-value for the mixed model correlation estimator can be determined with a normal approximation using the delta method (Lehmann, 1999). The correlation coefficient ρ_{XY} can be estimated from the covariance estimates. We know that

$$\rho_{UW} = \frac{\sigma_{UW}}{\sqrt{\sigma_U^2 \sigma_W^2}}$$

and since, for covariance parameters a ($cov_{ID}(1{,}1)$), b ($cov_{ID}(2{,}1)$), c ($cov_{ID}(2{,}2)$), g ($cov_{visit(ID)}(1{,}1)$), h ($cov_{visit(ID)}(2{,}1)$), i ($cov_{visit(ID)}(2{,}2)$), where $b + v = \sigma_{UW}$, $a + u = \sigma_U^2$, and $c + w = \sigma_W^2$, we have $\hat{\rho}_{UW} = \frac{b+h}{\sqrt{(a+g)*(c+i)}}$. We can estimate the standard error of ρ_{UW}, $\widehat{SE}(\rho_{UW})$, through derivation using the delta method. By the delta method, $\sqrt{n}(R_n - \rho_{UW}) \rightarrow N(0, \gamma^2)$ where $\gamma^2 = \left(\frac{\partial f}{\partial .}\right)' \boldsymbol{\Sigma} \left(\frac{\partial f}{\partial .}\right)$. We have the estimator R_n for ρ_{UW} where

$$R_n = \frac{cov_{ID}(2,1) + cov_{visit(ID)}(2,1)}{\sqrt{\left(cov_{ID}(1,1) + cov_{visit(ID)}(1,1)\right)\left(cov_{ID}(2,2) + cov_{visit(ID)}(2,2)\right)}} = \frac{b+h}{\sqrt{(a+g)*(c+i)}}$$

This estimator is made up of $Y_1^{(n)} = cov_{ID}(1,1) = a$, $Y_2^{(n)} = cov_{ID}(2,1) = b$, $Y_3^{(n)} = cov_{ID}(2,2) = c$, $Y_4^{(n)} = cov_{visit(ID)}(1,1) = g$, $Y_5^{(n)} = cov_{visit(ID)}(2,1) = h$, and $Y_6^{(n)} = cov_{visit(ID)}(2,2) = i$. The term $\frac{\partial f}{\partial .}$ is a column vector made up of the partial derivatives of f such that

$$\frac{\partial f}{\partial .} = \left(\frac{\partial f}{\partial a}\ \frac{\partial f}{\partial b}\ \frac{\partial f}{\partial c}\ \frac{\partial f}{\partial g}\ \frac{\partial f}{\partial h}\ \frac{\partial f}{\partial i}\right)'.$$

The partial derivatives of the function f are:

$$\frac{\partial f}{\partial a} = -\frac{1}{2}\frac{(b+h)(c+i)}{\sqrt{[(a+g)*(c+i)]^3}} \qquad \frac{\partial f}{\partial g} = -\frac{1}{2}\frac{(b+h)(c+i)}{\sqrt{[(a+g)*(c+i)]^3}}$$

$$\frac{\partial f}{\partial b} = \frac{1}{\sqrt{(a+g)*(c+i)}} \qquad \frac{\partial f}{\partial h} = \frac{1}{\sqrt{(a+g)*(c+i)}}$$

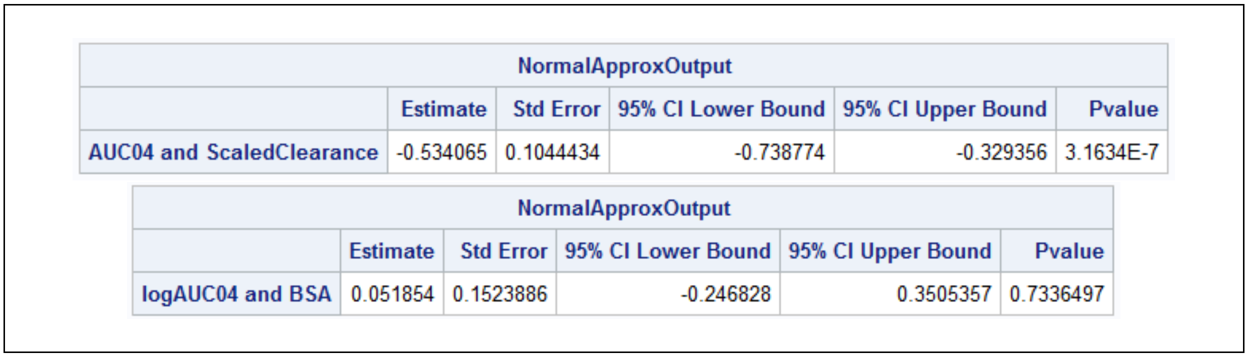
$$\frac{\partial f}{\partial c} = -\frac{1}{2}\frac{(b+h)(a+g)}{\sqrt{[(a+g)*(c+i)]^3}} \qquad \frac{\partial f}{\partial i} = -\frac{1}{2}\frac{(b+h)(a+g)}{\sqrt{[(a+g)*(c+i)]^3}}.$$

The matrix $\boldsymbol{\Sigma}$ is a 6x6 matrix made up of the components $\sigma_{11}, \sigma_{12}, \ldots, \sigma_{66}$. The component σ_{ij} is found as $cov(Y_i^n, Y_j^n)$. Thus $\sigma_{11} = var(a)$, $\sigma_{12} = cov(a,b)$, etc. From SAS, we can obtain the asymptotic covariance matrix of the covariance estimates using the ASYCOV option in PROC MIXED. These are used to determine the components of $\boldsymbol{\Sigma}$. With $\hat{\rho}_{UW}$ and $\widehat{SE}(\rho_{UW})$, we can make inferences on the correlation using the normal approximation (confidence interval and p-value) where the confidence interval is found as $\hat{\rho}_{UW} \pm 1.96\left(\widehat{SE}(\rho_{UW})\right)$ and the p-value is found using the cumulative distribution function of the normal distribution.

We fit the pharmacokinetics data using the SAS macro %MMCORR_NORMALAPPROX (Appendix I), which fits Hamlett's mixed model and use PROC IML to calculate the standard error using the derivation above. The macro call requires the inputs DATA (name of SAS dataset), ID (identifier for each subject), REP (the repeated measures identifier, such as visit number), VAR1 and VAR2 (variables to calculate correlation between). The macro calls to evaluate the 2 pairs of variables on the pharmacokinetics data are shown below:

```
%MMCorr_NormalApprox(data=pk,ID=ID,rep=visit,var1=AUC04,
             var2=ScaledClearance);
%MMCorr_NormalApprox(data=pk,ID=ID,rep=visit,var1=logAUC04,var2=BSA);
```

The SAS output for both macro calls is shown below:



NormalApproxOutput

	Estimate	Std Error	95% CI Lower Bound	95% CI Upper Bound	Pvalue
AUC04 and ScaledClearance	-0.534065	0.1044434	-0.738774	-0.329356	3.1634E-7

NormalApproxOutput

	Estimate	Std Error	95% CI Lower Bound	95% CI Upper Bound	Pvalue
logAUC04 and BSA	0.051854	0.1523886	-0.246828	0.3505357	0.7336497

Output 1. Output from %MMCORR_NORMALAPPROX

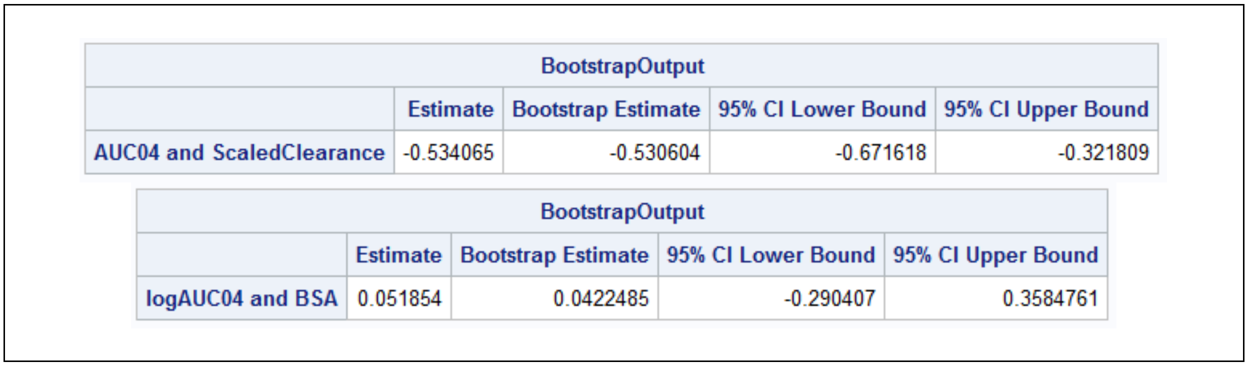
For AUC04 and Clearance/100, the confidence interval from the normal approximation is (-0.7388, -0.3294). The p-value is <0.0001, indicating that the correlation is significantly different than zero. In the case of log(AUC04) and BSA, the estimated confidence interval is (-0.2468, 0.3505) and the p-value is 0.7336. Both the confidence interval and p-value suggest that the correlation is not significantly different than zero.

CLUSTER BOOTSTRAP (CONFIDENCE INTERVAL)

The cluster bootstrap, described previously, is appropriate for calculating the confidence interval of the correlation coefficient from the mixed model approach. In this approach, the resampling occurs at the highest level and is evaluated using Hamlett et al.'s (2004) mixed model approach. The resampling and mixed model fitting is repeated B times, and the distribution of the estimates is used to form a confidence interval. The bias corrected and percentile intervals are reported for each combination of pharmacokinetics variables below. We resample 10,000 bootstrap samples for each pair of covariates using the SAS macro %MMCORR_BSCI (Appendix II). This macro requires similar inputs to the previous macro, with the additional variable NUMSAMPLES which specifies the number of bootstrap resamples:

```
%MMCorr_BSCI(data=pk,ID=ID,rep=visit,var1=AUC04,
       var2=ScaleClearance,numSamples=10000);
%MMCorr_BSCI(data=pk,ID=ID,rep=visit,var1=logAUC04,
       var2=BSA,numSamples=10000);
```

The output for both macro calls is shown below:



BootstrapOutput				
	Estimate	Bootstrap Estimate	95% CI Lower Bound	95% CI Upper Bound
AUC04 and ScaledClearance	-0.534065	-0.530604	-0.671618	-0.321809

BootstrapOutput				
	Estimate	Bootstrap Estimate	95% CI Lower Bound	95% CI Upper Bound
logAUC04 and BSA	0.051854	0.0422485	-0.290407	0.3584761

Output 2. Output from %MMCORR_BSCI

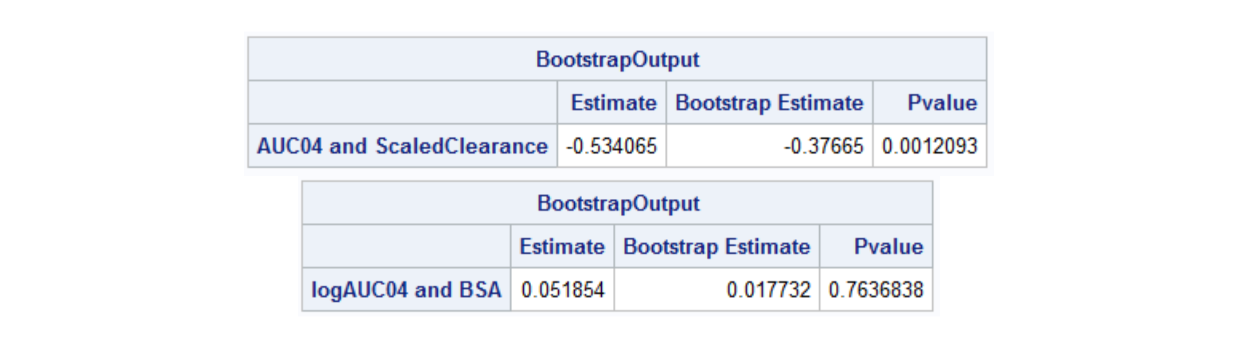
Using the cluster bootstrap approach, we evaluate the confidence interval for AUC04 and Clearance/100. The convergence rate was 99.78% and the bootstrap mean for the correlation coefficient is -0.5306. The cluster bootstrap bias corrected confidence interval is (-0.6716, -0.3218). This agrees with the previous finding that there is a statistically significant correlation between AUC04 and Clearance/100. Between log(AUC04) and BSA, the cluster bootstrap had a convergence rate of 99.55% and a bootstrap estimate of 0.0422. The bias corrected confidence interval is (-0.2904, 0.3585), which indicates that the correlation between these two covariates is not statistically significant.

CLUSTER BOOTSTRAP (P-VALUE)

The cluster bootstrap approach can be adapted to estimate the p-value for the correlation coefficient. In order to evaluate the p-value for bootstrap approaches, the bootstrap sampling scheme must be adjusted to reflect the null hypothesis ρ_{XY}=0. For the mixed model procedure, this can be performed by selecting a cluster boostrap sample (resample with replacement at the highest level). For the m_i observations within each cluster of the bootstrap sample, shuffle the Y_i's and rematch with X_i's to form m_i new pairs of observations. This reflects the null hypothesis as shuffling within a cluster should result in no correlation between the covariates within a subject. The mixed model can then be fit to each bootstrapped and shuffled sample to obtain the estimate $\hat{\rho}^{*(i)}$, which is repeated for B boostrap samples. The p-value can be calculated by determining the number of bootstrap estimates that are more extreme than the observed correlation estimate in the data. We implement cluster bootstrap with 10,000 resamples using shuffling to determine the p-value for both pairs of covariates. This can be calculated using the SAS macro %MMCORR_BSPVALUE (Appendix III), which requires the same inputs as previously described:

```
%MMCorr_BSpvalue(data=pk,ID=ID,rep=visit,var1=AUC04,
       var2=ScaleClearance,numSamples=10000);
%MMCorr_BSpvalue(data=pk,ID=ID,rep=visit,var1=logAUC04,
       var2=BSA,numSamples=10000);
```

The SAS output is shown below:



BootstrapOutput			
	Estimate	Bootstrap Estimate	Pvalue
AUC04 and ScaledClearance	-0.534065	-0.37665	0.0012093

BootstrapOutput			
	Estimate	Bootstrap Estimate	Pvalue
logAUC04 and BSA	0.051854	0.017732	0.7636838

Output 3. Output from %MMCORR_BSPVALUE

Between AUC04 and Clearance/100, the mixed model convergence rate was 99.23% and the p-value was estimated to be 0.0012. For log(AUC04) and BSA, the convergence rate was 99.57% and the p-value was 0.7637. We see that both results reflect the same conclusions as earlier findings using the normal approximation (p-values = <0.0001 and 0.7336, respectively) and the cluster bootstrap confidence interval approaches.

CONCLUSIONS

The correlation is a common measure of association between two variables. While it is standardized and simple to interpret, the correlation coefficient needs to account for multiple measurements per subject in repeated measures studies. Methods to estimate correlation in repeated measures data have been proposed, although a comprehensive comparative study has not been previously performed. We evaluated various approaches to calculate the correlation coefficient for repeated measures data and found that the mixed model approach proposed by Hamlett et al. (2004) is the most appropriate correlation measure. We recommend the mixed model approach as it explicitly defines the correlation structure within a subject's repeated measurements. We also introduce inference methods, including the normal approximation and cluster bootstrapping, to evaluate the strength of the correlation using confidence intervals and p-values.

REFERENCES

Bland, J.M. and Altman, D.G. 1994. “Correlation, regression, and repeated data.” British Medical Journal 308:896.

Christensen, R. 2011. Plane Answers to Complex Questions. New York, NY: Springer.

DiCiccio, T.J. and Efron, B. 1996. “Bootstrap Confidence Intervals.” Statistical Science 11(3):189-228.

Efron, B. 1979. “Bootstrap Methods: Another Look at the Jackknife.” The Annals of Statistics 7(1):1-26.

Field, C.A. and Welsh, A.H. 2007. “Bootstrapping Clustered Data.” Journal of the Royal Statistical Society Series B 69(3):369-390.

Hamlett, A., Ryan, L., and Wolfinger, R. 2004. “On the use of PROC MIXED to Estimate Correlation in the Presence of Repeated Measures.” SUGI 29 Proceedings.

Kim, H., Brooks, K.M., Tang, C.C., Wakim, P., Blake, M., Brooks, S.R., Montealegre Sanchez, G.A., de Jesus, A.A., Huang, Y., Tsai, W.L., Gadina, M., Prakash, A., Janes, J.M., Zhang, X., Macias, W.L., Kumar, P., and Goldbach-Mansky, R. 2017. “Pharmacokinetics, Pharmacodynamics, and Proposed Dosing of the Oral JAK1 and JAK3 Inhibitor Baricitinib in Pediatric and Young Adult CANDLE and SAVI Patients.” Clinical Pharmacology and Therapeutics doi: 10.1002/cpt.936.

Lehmann, E.L. 1999. Elements of Large-Sample Theory. New York, NY: Springer.

Lipsitz, S.R., Leong, T., Ibrahim, J., Lipshultz, S. 2001. “A Partial Correlation Coefficient and Coefficient of Determination for Multivariate Normal Repeated Measures Data.” Journal of the Royal Statistical Society Series D 50(1):87-95.

Roy, A. 2006. “Estimating Correlation Coefficient between Two Variables with Repeated Observations using Mixed Effects Model.” Biometrical Journal 48(2):286-301.

Shen, D., Lu, Z. 2006. “Computation of Correlation Coefficient and Its Confidence Interval in SAS.” SUGI 31 Proceedings.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Raphaela Goldbach-Mansky (NIAID, NIH) and Dr. Hanna Kim (NIAMS, NIH) for use of the pharmacokinetics data.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Katherine Irimata
Arizona State University
katherine.irimata@asu.edu

APPENDIX I: SAS MACRO FOR MIXED MODEL CORRELATION ESTIMATION USING THE NORMAL APPROXIMATION (CONFIDENCE INTERVAL AND P-VALUE)

```
%macro MMCorr_NormalApprox(data=,ID=,rep=,var1=,var2=);
data &data._long (drop=&var1. &var2. i); *univariate format;
      set &data. (keep=&ID. &rep. &var1. &var2.);
      array var[2] &var1. &var2.;
      do i = 1 to 2;
            Vtype = i;
            Response = var[i];
            output;
      end;
run;

/*Fit Hamlett (2004) mixed model*/
proc mixed data=&data._long method=ml covtest asycov asycorr;
      class &ID. vtype &rep.;
      model response = vtype / solution ddfm=kr;
      random vtype / type=un subject=&ID. v vcorr;
      repeated vtype / type=un subject=&rep.(&ID.);
      ods output VCorr=VCorr ConvergenceStatus=CS asycorr=asycorr
asycov=asycov CovParms=CovParms;
run;

/*Use delta method to find SE(rho), calculate CI/p-value based on normality*/
proc iml;
      use Covparms;
      read all var {Estimate} into Cov_estimate;
      close Covparms;

      use Asycov;
      read all var {CovP1 CovP2 CovP3 CovP4 CovP5 CovP6} into asycov;
      close Asycov;

      a = Cov_estimate[1];
      b = Cov_estimate[2];
      c = Cov_estimate[3];
      g = Cov_estimate[4];
      h = Cov_estimate[5];
      i = Cov_estimate[6];
      rhohat = (b+h)/sqrt((a+g)*(c+i));

      df_da = -0.5*((b+h)*(c+i))/sqrt(((a+g)*(c+i))**3);
      df_db = 1/sqrt((a+g)*(c+i));
      df_dc = -0.5*(b+h)*(a+g)/sqrt(((a+g)*(c+i))**3);
      df_dg = -0.5*(b+h)*(c+i)/sqrt(((a+g)*(c+i))**3);
      df_dh = 1/sqrt((a+g)*(c+i));
      df_di = -0.5*(b+h)*(a+g)/sqrt(((a+g)*(c+i))**3);

      create partialderiv var {df_da df_db df_dc df_dg df_dh df_di};
      append;
      close partialderiv;

      use partialderiv;
      read all var {df_da df_db df_dc df_dg df_dh df_di} into partialderiv;
      close partialderiv;
      Sigma = asycov;

      rho_var = partialderiv*Sigma*t(partialderiv);
      rho_SE = sqrt(rho_var);

      l95=rhohat-1.96*rho_SE;
      u95=rhohat+1.96*rho_SE;
      p = (1-cdf("normal",abs(rhohat),0,rho_SE))*2;

      NormalApproxOutput = j(1,5,.);
      NormalApproxOutput[,1]=rhohat;
      NormalApproxOutput[,2]=rho_SE;
      NormalApproxOutput[,3]=l95;
      NormalApproxOutput[,4]=u95;
      NormalApproxOutput[,5]=p;
      Outtitle={'Estimate'  'Std Error'  '95% CI Lower Bound'  '95% CI Upper
Bound' 'Pvalue'};
      Varnames=t("&var1. and &var2.");

      PRINT NormalApproxOutput[colname=Outtitle rowname=Varnames];

quit;
%mend MMCorr_NormalApprox;
```

APPENDIX II: SAS MACRO FOR MIXED MODEL CORRELATION ESTIMATION USING THE CLUSTER BOOTSTRAP (CONFIDENCE INTERVAL)

```
%macro MMCorr_BSCI(data=,ID=,rep=,var1=,var2=,numSamples=);
/*Select bootstrap samples*/
proc sort data=&data.(keep=&ID.) out=&data._ID nodupkey;
      by &ID.;
run;
proc surveyselect data=&data._ID NOPRINT seed=864132
out=&data._boot(drop=NumberHits)
      method=urs samprate=1 OUTHITS reps=&NumSamples.;
run;
data &data._boot;
      set &data._boot;
      if first.&ID. then RptNum = 1;
      else RptNum + 1;
      by Replicate &ID.;
      ID2 = &ID. || "_" || strip(RptNum);
run;
data &data._boot;
      set &data._boot;
      if first.Replicate then SubjNum = 1;
      else SubjNum + 1;
      by Replicate;
run;

data results;
      input Replicate rho12 convstatus reason $200. rho1 rho2;
run;

/*Bootstrap: Sample Individuals*/
%do j=1 %to &NumSamples.;
      proc sql;
      create table bootsamp
      as select a.*,b.&rep.,b.&var1.,b.&var2.
      from &data._boot a
      left join &data. b
      on &data._boot.ID = &data..ID
      where Replicate = &j.;
quit;
proc sort data=bootsamp;
      by &ID. RptNum &rep.;
run;

data &data._long (drop=&var1. &var2. i); *univariate format;
      set bootsamp (keep=&ID. RptNum ID2 &rep. &var1. &var2.);
      if first.ID2 then Visit2 = 1;
      else Visit2 + 1;
      by ID2;
      array var[2] &var1. &var2.;
      do i = 1 to 2;
            Vtype = i;
            Response = var[i];
            output;
            end;
run;

/*Fit Hamlett (2004) mixed model for bootstrap sample*/
proc mixed data=&data._long method=ml;
      class ID2 vtype visit2;
      model response = vtype / solution ddfm=kr;
      random vtype / type=un subject=ID2 v vcorr;
      repeated vtype / type=un subject=visit2(ID2);
      ods output VCorr=VCorr ConvergenceStatus=CS;
run;

/*Calculate correlation*/
proc iml;
      use Vcorr(keep=COL:);
      read all var _ALL_ into Vcorr;
      close Vcorr;

      use CS;
      read all var {Status} into convstatus;
      read all var {Reason} into reason;
      close CS;

      R = Vcorr[1:2,1:2];
      v = cusum( 1 || (ncol(R):2) );
      rho = remove(vech(R), v);
      Replicate = j(2*(2-1)/2,1,&j.);

      create corrcalc var {Replicate rho convstatus reason};
      append;
      close corrcalc;
quit;

data results;
      set results corrcalc;
run;
```

```
      dm "output;clear;log;clear;odsresults;select all;clear;";
%end;

data results2;
      set results;
      where convstatus = 0;
run;

data &data._long (drop=&var1. &var2. i); *univariate format;
      set &data. (keep=&ID. &rep. &var1. &var2.);
      array var[2] &var1. &var2.;
      do i = 1 to 2;
            Vtype = i;
            Response = var[i];
            output;
      end;
run;

/*Fit Hamlett (2004) mixed model for original dataset*/
proc mixed data=&data._long method=ml covtest asycov asycorr;
      class &ID. vtype &rep.;
      model response = vtype / solution ddfm=kr;
      random vtype / type=un subject=&ID. v vcorr;
      repeated vtype / type=un subject=&rep.(&ID.);
      ods output CovParms=CovParms;
run;

proc iml;
      use results2;
      read all var {rho} into rho;
      close results2;

      numsamp = nrow(rho);
      use Covparms;
      read all var {Estimate} into Cov_estimate;
      close Covparms;

      a = Cov_estimate[1];
      b = Cov_estimate[2];
      c = Cov_estimate[3];
      g = Cov_estimate[4];
      h = Cov_estimate[5];
      i = Cov_estimate[6];
      rhohat = (b+h)/sqrt((a+g)*(c+i));
      BS_mean = mean(rho);

      z = quantile("Normal",mean(rho < rhohat));
      a = 0;
      qlb = cdf("Normal",z+((z+quantile("Normal",0.025))/(1-
a*(z+quantile("Normal",0.025)))));
      qub = cdf("Normal",z+((z+quantile("Normal",0.975))/(1-
a*(z+quantile("Normal",0.975)))));

      call qntl(lb,rho,qlb);
      call qntl(ub,rho,qub);

      BootstrapOutput = j(1,4,.);
      BootstrapOutput[,1]=rhohat;
      BootstrapOutput[,2]=BS_mean;
      BootstrapOutput[,3]=lb;
      BootstrapOutput[,4]=ub;
      Outtitle={'Estimate'  'Bootstrap Estimate'  '95% CI Lower Bound'  '95%
CI Upper Bound'};
      Varnames=t("&var1. and &var2.");
      PRINT BootstrapOutput[colname=Outtitle rowname=Varnames];

quit;
%mend MMCorr_BSCI;
```

APPENDIX III: SAS MACRO FOR MIXED MODEL CORRELATION ESTIMATION USING THE CLUSTER BOOTSTRAP (P-VALUE)

```
%macro MMCorr_BSpvalue(data=,ID=,rep=,var1=,var2=,numSamples=);
/*Select bootstrap samples*/
proc sort data=&data.(keep=&ID.) out=&data._ID nodupkey;
      by &ID.;
run;
proc surveyselect data=&data._ID NOPRINT seed=864132
out=&data._boot(drop=NumberHits)
      method=urs samprate=1 OUTHITS reps=&NumSamples.;
run;
data &data._boot;
      set &data._boot;
      if first.&ID. then RptNum = 1;
      else RptNum + 1;
      by Replicate &ID.;
      ID2 = &ID. || "_" || strip(RptNum);
run;
data &data._boot;
      set &data._boot;
      if first.Replicate then SubjNum = 1;
      else SubjNum + 1;
      by Replicate;
run;

data results;
      input Replicate rho convstatus reason $200.;
run;

/*Bootstrap: Sample Individuals*/
%do j=1 %to &NumSamples.;
      proc sql;
            create table bootsamp
            as select a.*,b.&rep.,b.&var1.,b.&var2.
            from &data._boot a
            left join &data. b
            on &data._boot.ID = &data..ID
            where Replicate = &j.;
      quit;

      proc sort data=bootsamp;
            by &ID. RptNum &rep.;
      run;
      /*shuffle Wi's within subject, rematch with Ui's form m_i new pairs*/
      data w_order;
            set bootsamp (keep=ID2 SubjNum &rep. &var2.);
            order = ranuni(0);
      run;
      proc sort data=w_order;
            by ID2 order;
      run;

      /*create line number variable and merge on the line number*/
      data bootsamp;
            set bootsamp(drop=&var2.);
            linenum = _n_;
      run;
      data w_order;
            set w_order(keep=&var2.);
            linenum = _n_;
      run;
      data bootsamp;
            merge bootsamp w_order;
            by linenum;
      run;

      data &data._long (drop=&var1. &var2. i); *univariate format;
            set bootsamp (keep=&ID. RptNum ID2 &rep. &var1. &var2.);
            if first.ID2 then Visit2 = 1;
            else Visit2 + 1;
            by ID2;
            array var[2] &var1. &var2.;
            do i = 1 to 2;
                  Vtype = i;
                  Response = var[i];
                  output;
            end;
      run;

      /*Fit Hamlett (2004) mixed model for bootstrap sample*/
      proc mixed data=&data._long method=ml;
            class ID2 vtype visit2;
            model response = vtype / solution ddfm=kr;
            random vtype / type=un subject=ID2 v vcorr;
            repeated vtype / type=un subject=visit2(ID2);
            ods output VCorr=VCorr ConvergenceStatus=CS V=V
SolutionF=SolutionF;
      run;

      /*Calculate correlation for each bootstrap sample*/
      proc iml;
            use Vcorr(keep=COL:);
            read all var _ALL_ into Vcorr;
            close Vcorr;

            use V(keep=COL:);
            read all var _ALL_ into VCov;
            close V;

            use CS;
            read all var {Status} into convstatus;
            read all var {Reason} into reason;
            close CS;

            use SolutionF;
            read all var {Estimate} into mu;
            close SolutionF;

            R = Vcorr[1:2,1:2];
            v = cusum( 1 || (ncol(R):2) );
            rho = remove(vech(R), v);
            Replicate = j(2*(2-1)/2,1,&j.);

            create corrcalc var {Replicate rho convstatus reason};
            append;
            close corrcalc;
      quit;

      data results;
            set results corrcalc;
      run;
      dm "output;clear;log;clear;odsresults;select all;clear;";
%end;

data results2;
      set results;
      where convstatus = 0;
run;

data &data._long (drop=&var1. &var2. i); *univariate format;
      set &data. (keep=&ID. &rep. &var1. &var2.);
      array var[2] &var1. &var2.;
      do i = 1 to 2;
            Vtype = i;
            Response = var[i];
            output;
      end;
run;

/*Fit Hamlett (2004) mixed model for original dataset*/
proc mixed data=&data._long method=ml covtest asycov asycorr;
      class &ID. vtype &rep.;
      model response = vtype / solution ddfm=kr;
      random vtype / type=un subject=&ID. v vcorr;
      repeated vtype / type=un subject=&rep.(&ID.);
      ods output CovParms=CovParms;
run;

/*Calculate p-value*/
proc iml;
      use results2;
      read all var {rho} into rho;
      close results2;

      BS_mean = mean(rho);
      numsamp = nrow(rho);
      use Covparms;
      read all var {Estimate} into Cov_estimate;
      close Covparms;

      a = Cov_estimate[1];
      b = Cov_estimate[2];
      c = Cov_estimate[3];
      g = Cov_estimate[4];
      h = Cov_estimate[5];
      i = Cov_estimate[6];
      rhohat = (b+h)/sqrt((a+g)*(c+i));

      if rhohat >0 then p = (sum(rho < 0)+sum(rho > 2*rhohat))/numsamp;
      else p = (sum(rho > 0)+sum(rho < 2*rhohat))/numsamp;

      BootstrapOutput = j(1,3,.);
      BootstrapOutput[,1]=rhohat;
      BootstrapOutput[,2]=BS_mean;
      BootstrapOutput[,3]=p;
      Outtitle={'Estimate'  'Bootstrap Estimate'  'Pvalue'};
      Varnames=t("&var1. and &var2.");
      PRINT BootstrapOutput[colname=Outtitle rowname=Varnames];
quit;
%mend MMCorr_BSpvalue;
```