Technical Report 0901

Department of Biostatistics

State University of New York at Buffalo

A generalized linear model approach to estimating and testing the equality of conditional correlations

Gregory E. Wilding [1], Xueya Cai *† [2], and Alan Hutson [1]

March 24, 2009

[1] *Department of Biostatistics, the State University of New York at Buffalo. Buffalo, NY, 14214* [2] *Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, 46202*

SUMMARY

Methods of testing the equality of conditional correlation of bivariate data across a third variable of interest (covariate) have been studied [9, 10, 11, 12, 13, 14, 15, 16]. Most of these methods, such as the test based on Fisher z-transformation [15], the likelihood ratio tests, and the $C(\alpha)$ statistics [9], are limited to the case where a categorical covariate is considered. When the covariate is numeric, existing methods typically categorize data into groups based on percentiles of the covariate before estimating and comparing sub-group correlations. As an example, consider a study where the Pearson correlation coefficient of diastolic blood pressure with weight is tested among people of different ages. Using current methods, an analyst would first define meaningful age groups before performing the equality test.

In this study, we propose a generalized linear model approach for estimation and hypothesis testing about the Pearson correlation coefficient, where the correlation itself can be modeled as a function of numeric covariates. This approach allows for flexible and robust inference and prediction of the

*Correspondence to: Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, 46202.

†E-mail: xueycai@iupui.edu

conditional correlations based on the generalized linear model. Simulation studies show that the proposed method is statistically more powerful and more flexible in accommodating complex covariate patterns than existing methods.

[keywords]: Correlation coefficient; equality test; generalized linear model; MOS SF-36

1. Introduction

The correlation coefficient, ρ, is a parameter that measures the linear association between two random variables. Different methods have been studied to estimate a common population correlation, and to test the hypothesis $H_0 : \rho = 0$ [1, 2, 3, 4, 5, 6, 7, 8]. In addition, approaches are available to testing the equality of correlations in bivariate populations across a third variable of interest (covariate) [9, 10, 11, 12, 13, 14, 15, 16], where the covariate is categorical.

When the covariate is categorical, or treated as categorical, let $(X_{ij}, Y_{ij}), i = 1, 2, \ldots, k, j = 1, 2, \ldots, n_i$ be k independent random samples from k bivariate normal distributions with means (μ_{X_i}, μ_{Y_i}), variances $(\sigma^2_{X_i}, \sigma^2_{Y_i})$, and correlations ρ_i. Current practices test the following hypothesis:

$$\begin{aligned} H_0 &: \rho_1 = \rho_2 = \ldots = \rho_k \\ H_1 &: \text{Not all the } \rho_i' s \text{ are equal.} \end{aligned} \quad (1)$$

For example, Steel and Torrie [15] propose a test assuming asymptotic normality of the Fisher z-transformation [5] of the sample correlation coefficients (χ^2_Z). Paul [9] compares the χ^2_Z test to several tests derived from a likelihood ratio statistic (χ^2_L), a $C(\alpha)$ test statistic based on Fisher-z transformation (χ^2_{CF}), and a $C(\alpha)$ statistic based on the maximum likelihood estimator of the common correlation coefficient (χ^2_{CM}). Paul recommends the χ^2_{CF} test for general use because of its computational ease and slight power advance. In addition, Wilcox [13] implements a method based on percentile bootstrap simulation for the comparison of correlation coefficients when the conditional variance of variable Y, given X, is not a constant.

Finally, Yu and Dunn [14] have studied nonparametric methods, such as the "Spearman" test and the "Kendall" test, and found that they perform better than the statistics based on Fisher z-transformation for non-normal data.

Few studies have considered the situation where the covariate of interest is numeric, or where a continuous rather than discrete change of the correlation coefficient is being conceived. One exception may be the approach by Bartlett [16]. In Bartlett's paper, the author considers a linear model for the Pearson's correlation coefficient over a numeric grouping variable, in which the correlation coefficient at each level of a categorized version of the third variable is first estimated. Weighted least square estimators are obtained, and their asymptotic distributions are discussed. This regression based approach may represent an improvement over previous methods. However, its limitations are that an overall large sample size is necessary for estimating the association between the Pearson correlation coefficient and the covariate, and the model becomes quickly intractable when two or more covariates of interest exist. Therefore, the approach still cannot fully accommodate numeric covariate(s).

This study proposes a more flexible methodology for modeling the Pearson correlation coefficient of bivariate normal data across both numeric and categorical covariates. The outline of this paper is as follows. Section 2 introduces the generalized linear model (GLM) approach to estimating the correlation coefficient for bivariate normal data, where the restricted maximum likelihood (REML) approach is used for parameter estimation. Simulation studies are performed in Section 3 to compare the type I error rate and power of the proposed method to those of existing methods, and to test the robuestness of the proposed approach when the correlation function or data distribution is misspecified. Section 4 presents a real data example. The strength and limitation of the proposed method is discussed in Section 5.

2. The Proposed Method

We assume X_i and Y_i to be bivariate normally distributed, where their means are considered functions of the covariates $\boldsymbol{Z_{X_i}}$ and $\boldsymbol{Z_{Y_i}}$, respectively, and their variances σ_X^2 and σ_Y^2 are constants. The data distribution is expressed as (in a matrix form)

Definition 2.1. *Let*

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \Big| \boldsymbol{Z} \sim MVN\left(\begin{bmatrix} \boldsymbol{\mu_{X|Z}} \\ \boldsymbol{\mu_{Y|Z}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_{XX}} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_{YY}} \end{bmatrix} \right),$$

where

$$\begin{aligned} \boldsymbol{X'} &= (x_1, x_2, \ldots, x_n), \\ \boldsymbol{Y'} &= (y_1, y_2, \ldots, y_n), \\ \boldsymbol{\mu_{X|Z}} &= \boldsymbol{Z_X \beta_X}, \\ \boldsymbol{\mu_{Y|Z}} &= \boldsymbol{Z_Y \beta_Y}, \\ \boldsymbol{\Sigma_{XX}} &= \sigma_X^2 \boldsymbol{I}, \\ \boldsymbol{\Sigma_{YY}} &= \sigma_Y^2 \boldsymbol{I}, \\ \boldsymbol{\Sigma_{XY}} &= \ \boldsymbol{\Sigma_{YX}} = \boldsymbol{\Sigma_{XX}^{1/2}} \boldsymbol{Diag}(\rho_i) \boldsymbol{\Sigma_{YY}^{1/2}} \end{aligned}$$

where $\boldsymbol{\beta_X}$ and $\boldsymbol{\beta_Y}$ are parameter vectors in the regression models of X on Z and Y on Z, $\boldsymbol{Z_X}$ and $\boldsymbol{Z_Y}$ are design matrices corresponding to the regression models of X on Z and Y on Z, respectively, $\boldsymbol{I}$ is identity matrix, and ρ_i is the Pearson correlation coefficient between X and Y that varies over Z. $\boldsymbol{Z_X}$ and $\boldsymbol{Z_Y}$ need not be the same.

We further define a function $h^{-1}(\cdot)$ that relates ρ_i to $\boldsymbol{Z_{\gamma_i}}$, where $\boldsymbol{Z_{\gamma_i}}$ is the i^{th} row of the design matrix $\boldsymbol{Z_\gamma}$ corresponding to the correlation vector $\boldsymbol{\rho}$. Here $\boldsymbol{Z_\gamma}$ need not be the same as $\boldsymbol{Z_X}$ or $\boldsymbol{Z_Y}$. Since $(\rho_i + 1)/2$ has a range of [0,1], one form of the link function could be

$$h(\rho_i \mid Z) = \Phi^{-1}\left(\frac{\rho_i + 1}{2}\right) = \gamma_0 + \gamma_1 z_i \tag{2}$$

when only one covariate is considered in the function, where $\Phi(\cdot)$ represents the CDF of the standard normal random variable, and γ_0 and γ_1 are covariate parameters. Equation (2) can be re-written as

$$\rho_i = 2\Phi\left(\gamma_0 + \gamma_1 z_i\right) - 1. \tag{3}$$

Therefore the log likelihood can be written as

$$\begin{aligned}
& l(\mu_{X_i}, \mu_{Y_i}, \sigma_X^2, \sigma_Y^2, \rho_i | X, Y, Z) \\
= \quad & -n\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log\left(\sigma_X^2\right) - \frac{1}{2}\sum_{i=1}^{n}\log(\sigma_Y^2) - \frac{1}{2}\sum_{i=1}^{n}\log\left(1-\rho_i^2\right) \\
& -\sum_{i=1}^{n}\frac{(x_i-\mu_{X_i})^2}{2(1-\rho_i^2)\sigma_X^2} + \sum_{i=1}^{n}\frac{\rho_i(x_i-\mu_{X_i})(y_i-\mu_{Y_i})}{\sigma_X\sigma_Y(1-\rho_i^2)} - \sum_{i=1}^{n}\frac{(y_i-\mu_{Y_i})^2}{2(1-\rho_i^2)\sigma_Y^2},
\end{aligned} \tag{4}$$

where μ_{X_i}, μ_{Y_i} and ρ_i can be replaced by the link functions denoted above.

Parameters in the mean functions are not of interest in our study, and thus can be considered nuisance parameters. The restricted maximum likelihood (REML) approach can be used in order to eliminate these nuisance parameters[17].

The following theorem is needed to apply the REML approach to our proposed method.

Theorem 2.2. *Let*

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}$$

be given by Definition 2.1. Also let

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{X}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Y}} \end{bmatrix},$$

where $\boldsymbol{H}_{\boldsymbol{X}}$ and $\boldsymbol{H}_{\boldsymbol{Y}}$ are hat matrices of the design matrices of $\boldsymbol{Z}_{\boldsymbol{X}}$ and $\boldsymbol{Z}_{\boldsymbol{Y}}$, respectively:

$$\begin{aligned}
\boldsymbol{H}_{\boldsymbol{X}} &= \boldsymbol{Z}_{\boldsymbol{X}}(\boldsymbol{Z}_{\boldsymbol{X}}'\boldsymbol{Z}_{\boldsymbol{X}})^{-1}\boldsymbol{Z}_{\boldsymbol{X}}', \\
\boldsymbol{H}_{\boldsymbol{Y}} &= \boldsymbol{Z}_{\boldsymbol{Y}}(\boldsymbol{Z}_{\boldsymbol{Y}}'\boldsymbol{Z}_{\boldsymbol{Y}})^{-1}\boldsymbol{Z}_{\boldsymbol{Y}}'.
\end{aligned}$$

Now let

$$\boldsymbol{W}^{*} = \begin{bmatrix} \boldsymbol{X}^{*} \\ \boldsymbol{Y}^{*} \end{bmatrix} = \boldsymbol{C} \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix},$$

then the conditional covariance matrix is

$$\boldsymbol{\Sigma}_{\boldsymbol{X}^*\boldsymbol{Y}^*} = \sqrt{\boldsymbol{\Sigma}_{\boldsymbol{X}^*\boldsymbol{X}^*}}\,\boldsymbol{Diag}(\rho_i)\sqrt{\boldsymbol{\Sigma}_{\boldsymbol{Y}^*\boldsymbol{Y}^*}},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{X}^*\boldsymbol{X}^*}$ *and* $\boldsymbol{\Sigma}_{\boldsymbol{Y}^*\boldsymbol{Y}^*}$ *are the variance matrices for* $\boldsymbol{X}^*$ *and* $\boldsymbol{Y}^*$*, respectively. That is, the conditional correlation matrix of the transformed data* W^* *equals the one of the original data, and*

$$E\left(\boldsymbol{W}^{*}\right) = E \begin{bmatrix} \boldsymbol{X}^{*} \\ \boldsymbol{Y}^{*} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

The proof of Theorem 2.2 is shown in the Appendix. Given the Theorem, the transformed data remain the same conditional correlation structure as the original data, and nuisance parameters in the mean functions are eliminated. After data transformation, parameter estimates and their variance-covariance structure in the link function of ρ_i can be obtained using iteration techniques. The corrected likelihood ratio test (LRT_F) can be used for hypothesis testing, where the statistic is a function of the log likelihood ratio (λ):

$$U = \lambda^{2/n}, \tag{5}$$

and

$$\frac{1-U}{U}\frac{n-q}{q_1} \overset{Asm}{\sim} F\left(q_1, n-q\right), \tag{6}$$

where n is the number of observations, q is the number of parameters in the model, and q_1 is the number of parameters involved in the null hypothesis. A merit of the corrected likelihood ratio test is that the limiting distribution of the test statistic converges faster than the one based on -2λ [18, 19].

3. Simulation Study

In this section we perform Monte Carlo simulations based on random samples generated from the bivariate normal distribution

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \sim \mathrm{BVN}\left(\begin{bmatrix} 1+2Z_i \\ 1.5+1.2Z_i \end{bmatrix}, \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}\right), \tag{7}$$

where Z_i is randomly chosen from uniform(0,1) distribution. We further define the underlying link function of the correlation ρ_i aganist Z_i as follows:

$$\rho_i = 2\Phi\left(\gamma_0 + \gamma_1 Z_i\right) - 1 \tag{8}$$

After data are generated, parameters of the proposed GLM approach are estimated using the REML approach. Hypothesis testing for $\gamma_1 = 0$ is conducted with the LRT_F test described before. We compare the type I error rate and power of our approach to those of existing k-sample test methods [9, 15], where we first divide the simulated data into two or three groups based on the percentiles of $z_i, i = 1, 2, \ldots, n$, and then calculate the χ^2_z and χ^2_{CF} statistics for comparison purposes. We choose the two statistics for comparison because χ^2_Z is most commonly used, and χ^2_{CF} is recommended by Paul as an easily applied approach with potentially higher power compared to competitors [9].

Table I shows that both the proposed GLM method with LRT_F test and the two current methods maintain the type I error rate very well. Table II shows that the power of these tests increases with sample size and is dependent on the assumed correlation structures as would be expected. When the correlation between X and Y does not vary substantially across the covariate Z (i.e., $\gamma_1 = 0, 0.1, 0.2$), the LRT$_F$, χ^2_Z, and χ^2_{CF} tests have similarly low power, that is, for the different sample sizes considered in the simulation (n=30, 60, 90, 120, 150, 180), the power of all approaches is generally below 10%. However, when the correlation between X and Y varies in a more substantial way (i.e., $\gamma_1 = 0.5, 1.0, 1.5$), the LRT$_F$ test based on the proposed method is more powerful than existing methods (χ^2_Z and χ^2_{CF}).

Regarding the robustness of the proposed method, we test the equality of conditional correlations across the covariate when the true data are generated from the probit link function in Monte Carlo simulation (see equation (3)), while the estimated model defines the link function $h^{-1}(\cdot)$ based on Fisher z-transformation:

$$\rho_i = \frac{2\exp(\gamma_0 + \gamma_1 Z_i) - 1}{2\exp(\gamma_0 + \gamma_1 Z_i) + 1} \tag{9}$$

We also study the robustness of the proposed model (with normality assumption) when the true data are generated in Monte Carlo simulation as elliptical contoured distributed (specifically, Pearson Type II distributed). Both robustness tests are in comparison with the χ^2_Z and χ^2_{CF} tests.

Table III shows the estimated type I error rate and power of the three tests when the link function of ρ_i is misspecified. The three tests, even with misspecified link function based on Fisher z-transformation, maintain the type I error rate reasonably well. However, the power of the LRT$_F$ test based on the proposed GLM approach is higher than the power of two existing test statistics in almost all considered scenarios. The increased power of the proposed LRT$_F$ test is more pronounced when sample size or γ_1 increases. In general, even when the estimated correlation function is misspecified, the LRT$_F$ test based on the proposed method shows reasonably high power (i.e., >70%) when sample size is reasonably large ($n \geq 60$) and the true variation of ρ_i across the covariate is medium (i.e., $\gamma_1 \geq 0.5$).

Table IV compares the type I error rate and power of the proposed and existing methods with bivariate normal assumption while the true data are generated from Pearson type II distribution. The results suggest that all tests are conservative in maintaining the type I error rate. In many simulated scenarios, however, the LRT_F test based on the proposed method still shows higher power than the χ^2_Z and χ^2_{CF} tests.

4. An example

In this section, we present a real sample analysis based on published data from the 2003 and 2004 Health Outcome Survey (HOS) study [20], where we model the Pearson correlation coefficient between the physical component summary (PCS) and mental component summary (MCS) of the Medical Outcome Study 36-items Short Form Health Survey (MOS SF-36) as a function of demographic covariates of survey respondents. The MOS SF-36 was developed in 1990's to measure generic health status and quality of life [21], of which the PCS and MCS are aggregated components to quantify overall physical and mental health functions, respectively [22]. The medical literature has suggested that the two components may be correlated with age, and with each other [21, 23, 24, 25].

The data we use include 28,724 survey respondents sampled from 53 Medicare insurance plans [20]. Bivariate analysis shows positive correlation between PCS and MCS ($r = 0.14, P < 0.0001$). In addition, both of them are negatively correlated with age ($r = -0.22, P < 0.0001,$ and $r = -0.08, P < 0.0001$, respectively). we categorize age into twenty groups based on percentiles and calculate the Pearson correlation coefficient between PCS and MCS within each age group. Results suggest a nonlinear downward trend of the correlation coefficient over age (Figure 1).

To illustrate the proposed GLM approach, we model the correlation of PCS with MCS using a probit link function of age and other demographic variables, including marital status, gender,

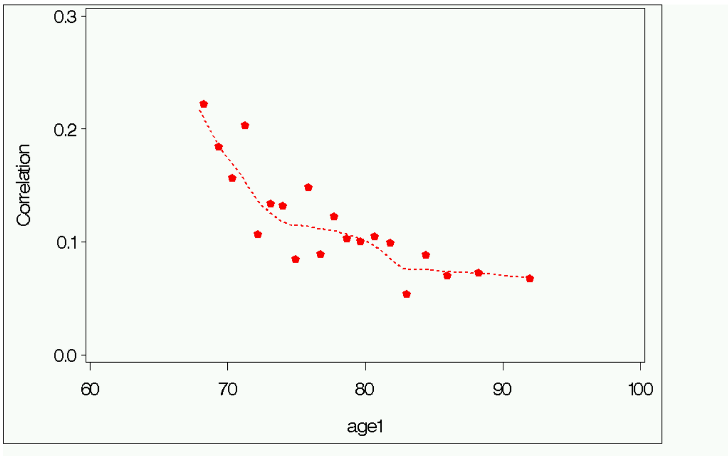

Figure 1. Correlation of PCS with MCS in the Health Outcome Survey data within twenty age groups

and race. That is,

$$\rho = 2\Phi\left(\gamma_0 + \gamma_1 \times \text{AGE} + \gamma_2 \times \text{AGE}^2 + \gamma_3 \times \text{AGE}^3 + \gamma_4 \times \text{MARRIED} + \gamma_5 \times \text{GENDER} + \gamma_6 \times \text{RACE}\right) - 1$$

where, based on the Akaike Information Criterion (AIC) for model selection [27, 28], the second and third order polynomials of age (in years) are included to model nonlinear trend, and other demographic variables are defined as below:

- MARRIED: 1 if married, 0 otherwise
- GENDER: 1 if female, 0 otherwise
- RACE: 1 if non-white, 0 white

The restricted maximum likelihood estimates of parameters are shown in Table V, and the estimated correlation function is

$$\rho = 2\Phi\left(0.194 - 0.024 \times \text{AGE} + 0.027 \times \text{AGE}^2 - 0.010 \times \text{AGE}^3 + 0.041 \times \text{MARRIED} - 0.124 \times \text{GENDER} - 0.020 \times \text{RACE}\right) - 1$$

It can be seen that age still significantly predicts the Pearson correlation coefficient between PCS and MCS even after controlling for other demographic variables. In addition, married and male respondents tend to show higher correlation between their physical and mental health functions than others, while the correlation does not seem to vary across racial groups (Table V). Figure 2 shows the predicted correlation between PCS and MCS as a function of age while stratified by marital status (panel 1) and gender (panel 2).

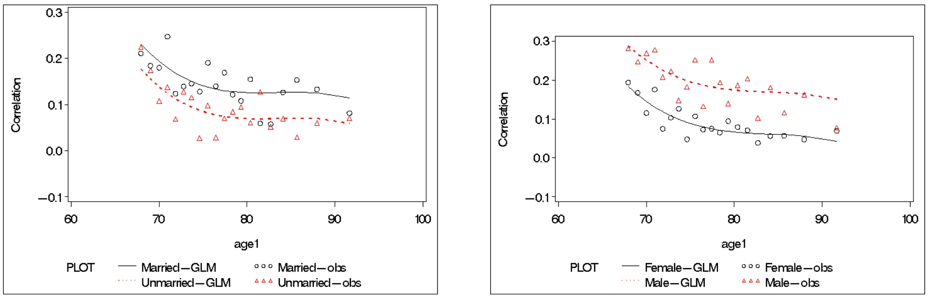

Figure 2. Predicted correlation between PCS and MCS over age in the HOS data: by marital status (Panel 1) and gender (Panel 2)

5. Discussion

Methods for testing the equality of linear association between two random variables across a categorical covariate have been studied in the literature [9, 15, 16]. In the presence of a numeric covariate, current practices are to first categorize the covariate into groups before calculating and comparing correlation coefficients of the two random variables across grouped samples. These approaches require relatively large sample so that sub-group correlation coefficients can be tested. In addition, categorization of data results in information loss and is not feasible when multiple covariates are of interest or need to be controlled for.

In this study, we propose a generalized linear model approach to estimating the Pearson correlation coefficient of bivariate normal variables and testing its equivalence across numeric covariate(s). This approach directly estimates the correlation coefficient as a function of numeric or categorical covariate(s), and allows for flexible and robust parameter estimation and inference through the REML approach. The proposed method maintains type I error rate well, and shows higher statistical power than existing approaches. We note that in the simplest case where the conditional correlation coefficient varies over a single categorical covariate, our proposed approach reduces to many existing approaches.

This proposed method has limitations. First, it is computationally more intensive than existing methods since it requires optimization technique for parameter estimation. Additionally, when the numeric covariate is of extreme values, the model with linear specification may lead to unreliable parameter estimates [26]. Future work is needed to improve the proposed approach to accommodate non-normal data or data that are obtained from non-independent (or clustered) samples, such as the case where the PCS and MCS scores are collected from respondents in (or potentially clustered by) multiple insurance plans.

ACKNOWLEDGEMENT

We would like to thank Dr. Yue Li at the University of California, Irvine for providing the HOS data.

APPENDIX

Proof of Theorem 2.2:

According to the properties of multivariate normal distribution [29], the transformed vector,

$$\mathbf{W}^* = \mathbf{C}\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{I} - \boldsymbol{H_X})\boldsymbol{X} \\ (\boldsymbol{I} - \boldsymbol{H_Y})\boldsymbol{Y} \end{bmatrix}$$

is multivariate normal distributed with expectation

$$\begin{bmatrix} \boldsymbol{I} - \boldsymbol{H_X} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} - \boldsymbol{H_Y} \end{bmatrix}\begin{bmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Y} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

and its covariance structure

$$\begin{aligned}&\begin{bmatrix} \boldsymbol{I}-\boldsymbol{H_X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}-\boldsymbol{H_Y} \end{bmatrix}\begin{bmatrix} \boldsymbol{\Sigma_{XX}} & \boldsymbol{\Sigma_{XY}} \\ \boldsymbol{\Sigma_{XY}} & \boldsymbol{\Sigma_{YY}} \end{bmatrix}\begin{bmatrix} \boldsymbol{I}-\boldsymbol{H_X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}-\boldsymbol{H_Y} \end{bmatrix} \\ &= \begin{bmatrix} (\boldsymbol{I}-\boldsymbol{H_X})\boldsymbol{\Sigma_{XX}}(\boldsymbol{I}-\boldsymbol{H_X}) & (\boldsymbol{I}-\boldsymbol{H_X})\boldsymbol{\Sigma_{XY}}(\boldsymbol{I}-\boldsymbol{H_Y}) \\ (\boldsymbol{I}-\boldsymbol{H_X})\boldsymbol{\Sigma_{XY}}(\boldsymbol{I}-\boldsymbol{H_Y}) & (\boldsymbol{I}-\boldsymbol{H_Y})\boldsymbol{\Sigma_{YY}}(\boldsymbol{I}-\boldsymbol{H_Y}) \end{bmatrix}.\end{aligned}$$

Since the matrices $\boldsymbol{I}-\boldsymbol{H_X}$ and $\boldsymbol{I}-\boldsymbol{H_Y}$ are idempotent symmetric [29], we have

$$\begin{aligned}\boldsymbol{\Sigma_{X^*X^*}} &= (\boldsymbol{I}-\boldsymbol{H_X})\boldsymbol{\Sigma_{XX}}(\boldsymbol{I}-\boldsymbol{H_X}) \\ &= (\boldsymbol{I}-\boldsymbol{H_X})\sigma_X^2\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{H_X}) \\ &= \sigma_X^2\left[(\boldsymbol{I}-\boldsymbol{H_X})\right]^2 \\ &= \left[\sigma_X(\boldsymbol{I}-\boldsymbol{H_X})\right]^2,\end{aligned}$$

and

$$\boldsymbol{\Sigma_{Y^*Y^*}} = (\boldsymbol{I}-\boldsymbol{H_Y})\boldsymbol{\Sigma_{YY}}(\boldsymbol{I}-\boldsymbol{H_Y}) = \left[\sigma_Y(\boldsymbol{I}-\boldsymbol{H_Y})\right]^2,$$

where $\boldsymbol{\Sigma_{X^*X^*}}$ and $\boldsymbol{\Sigma_{Y^*Y^*}}$ are the variance matrices of transformed X and Y, respectively. Therefore

$$\begin{aligned}\boldsymbol{\Sigma_{X^*Y^*}} &= (\boldsymbol{I}-\boldsymbol{H_X})\boldsymbol{\Sigma_{XY}}(\boldsymbol{I}-\boldsymbol{H_Y}) \\ &= (\boldsymbol{I}-\boldsymbol{H_X})\sigma_X\boldsymbol{I}\mathbf{Diag}(\boldsymbol{\rho_i})\sigma_Y\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{H_Y}) \\ &= \left[\sigma_X(\boldsymbol{I}-\boldsymbol{H_X})\right]\mathbf{Diag}(\boldsymbol{\rho_i})\left[\sigma_Y(\boldsymbol{I}-\boldsymbol{H_Y})\right] \\ &= \sqrt{\boldsymbol{\Sigma_{X^*X^*}}}\mathbf{Diag}(\rho_i)\sqrt{\boldsymbol{\Sigma_{Y^*Y^*}}} \qquad q.e.d.\end{aligned}$$

REFERENCES

1. Guo JH. Four correlation coefficients with a third blocking variable: their efficacy, relative efficiency, and test statistics. *Communications in Statistics: Theory and Methods* 2003 **32**(9): 1835-1858
2. Paul SR, Barnwal PK. Maximum likelihood estimation and a C(α) test for a common intraclass correlation. *The statistician* 1990 **39** (1): 19-24

3. Donner A, Rosner B. On inferences concerning a common correlation coefficient. *Applied Statistics* 1980; **29**(1): 69-76
4. Duncan GT, Layard MW. A Monte-Carlo study of asymptotically robust tests for correlation. *Biometrika* 1973; **60**: 551-558
5. Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 1915; **10**: 507-521
6. Fisher RA. On the probable error of a coefficient of correlation deduced from a small sample. *Metron* 1921; **1**: 3-32
7. Pourahmadi M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** (3): 677-690
8. Tian L, Wilding GE. Confidence interval estimation of a common correlation coefficient. *Computational Statistics and Data Analysis* 2008 **52**:4872-4877
9. Paul SR. Test for the equality of several correlation coefficients. *The Canadian Journal of Statistics* 1989; **17**(2): 217-227
10. Sakaori F. A nonparametric test for the equality of dependent correlation coefficients under normality. *Communications in Statistics: Theory and Methods* 2002 **31** (12): 2379-2389
11. Sakaori F. Permutation test for equality of correlation coefficients in two populations. *Communications in Statistics: Simulation* 2002 **31** (4): 641-651
12. Wolfe DA. A distribution-free test for related correlation coefficients. *Technometrics* 1977 **19** (4): 507-509
13. Wilcox RR, Muska J. Comparing correlation coefficients. *Communications in Statistics: Simulation* 2002 **31** (1): 49-59
14. Yu MC, Dunn OJ. Robust tests for the equality of two correlation coefficients: A Monte Carlo study. *Educational and Psychological Measurement* 1982; **42**: 987-1004
15. Steel RGD, Torrie JS. *Principles and Procedures of Statistics: A Biometrical Approach.* McGraw-Hill, 1980
16. Bartlett RF. Linear modeling of Pearson's product moment correlation coefficient: An application of Fisher's z-transformation. *The Statistician* 1993; **42**:45-53
17. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall, 1989
18. Anderson TW. *An Introduction to Multivariate Statistical Analysis*(3rd edn). Wiley, 2003.
19. Rao CR. An asymptotic expansion of the distribution of Wilks's criterion. *Bulletin of the international statistical institute* 1951; **33**(2): 177-180
20. Li Y, Cai X, Glance LG, Mukamel DB. Gender differences in healthcare-seeking behavior for urinary incontinence and the impact of socioeconomic status: a study of the medicare managed care population. *Medical Care* 2007; **45**: 1116-1122
21. Heyes V, Morris J, Wolfe C, Morgan M. The SF-36 health survey questionnaire - Is it suitable for use with older adults? *Age and Ageing* 1995; **24**: 120-125
22. Bierman AS, Lawrence WF, Haffer SC, Clancy CM. Functional health outcomes as a measure of health care quality for Medicare beneficiaries. *Health Services Research* 2001; **36**: 90-109

23. Mittal SK, Ahern L, Flaster E, et. al. Self-assessed physical and mental function of haemodialysis patients. *Nephrol Dial Transplant* 2001; **16**: 1387-1394

24. Iliescu EA, Coo H, McMurray MH, et. al. Quality of sleep and health-related quality of life in haemodialysis patients. *Nephrol Dial Transplant* 2003; **18**: 126-132

25. Newton-Bishop JA, Nolan C, Turner F, et. al. A quality-of-life study in high-risk (ThicknessX2 mm) cutaneous melanoma patients in a randomized trial of 1-cm versus 3-cm surgical excision margins. *Journal of Investigative Dermatology Symposium Proceedings* 2004; **9**: 152-159

26. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. *Applied Linear Statistical Models.* McGraw-Hill, 1996

27. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**(6): 716-723

28. Burnham KP, Anderson DR. *Model Selection and Inference: A Practical Information-Theoretic Approach.* Springer - Verlag, 1998.

29. Graybill FA. *Theory and Application of the Linear Model*. Duxbury Press, 2000

Table I. Estimated type I error rate (%) of tests based on the proposed (LRT_F) and existing (χ^2_Z and χ^2_{CF}) approaches for the equality test of conditional correlations ($\alpha = 0.05, \gamma_1 = 0$)

n	γ_0	ρ	Estimated α				
				n=(1:1)		n=(1:1:1)	
			LRT_F	χ^2_Z	χ^2_{CF}	χ^2_Z	χ^2_{CF}
30	0	0	6.0	4.7	6.6	5.3	6.8
	0.3	0.39	6.0	4.7	6.7	5.3	7.5
	0.6	0.84	5.3	4.8	6.2	5.2	7.2
60	0	0	5.4	5.3	5.7	5.4	6.4
	0.3	0.39	5.5	5.3	5.3	5.1	6.1
	0.6	0.84	5.0	5.0	6.1	5.0	5.7
90	0	0	5.6	5.1	5.8	5.5	5.5
	0.3	0.39	5.2	5.1	5.4	4.8	5.6
	0.6	0.84	4.9	4.9	5.5	4.7	5.6
120	0	0	5.2	5.0	5.7	4.8	5.4
	0.3	0.39	5.1	5.0	5.4	5.3	5.7
	0.6	0.84	4.9	4.9	5.0	4.8	5.7
150	0	0	5.1	4.9	5.5	5.0	5.1
	0.3	0.39	5.5	4.9	5.0	4.7	5.7
	0.6	0.84	5.3	4.6	5.5	5.0	5.3
180	0	0	5.3	5.1	5.6	5.3	5.5
	0.3	0.39	4.8	5.2	4.8	5.2	5.2
	0.6	0.84	5.4	5.1	5.2	5.1	4.9

Table II. Type I error rate and power (%) of the proposed (LRT_F) and existing (χ^2_Z and χ^2_{CF}) approaches for the equality test of conditional correlations ($\gamma_0 = 0.5$)

n	γ_1	Probability of Rejection				
			n=1:1		n=1:1:1	
		LRT_F	χ^2_Z	χ^2_{CF}	χ^2_Z	χ^2_{CF}
30	0	6.1	5.0	6.1	5.2	7.1
	0.1	6.0	4.6	6.6	5.0	6.8
	0.2	6.4	5.3	7.7	4.9	7.9
	0.5	10.6	8.2	9.2	7.7	8.7
	1	31.5	18.9	28.1	13.7	19.8
	1.5	56.0	49.8	37.1	27.9	41.2
60	0	5.2	5.2	5.5	5.1	5.8
	0.1	5.9	5.4	5.6	5.7	6.5
	0.2	6.9	6.3	6.9	10.1	6.8
	0.5	17.3	12.3	12.2	12.5	13.4
	1	56.2	29.6	42.7	36.9	40.9
	1.5	94.4	62.9	77.1	72.3	72.2
90	0	5.3	5.0	5.7	4.9	5.8
	0.1	6.2	5.5	6.0	5.4	5.9
	0.2	8.3	6.7	7.4	5.8	6.9
	0.5	25.2	18.7	18.6	11.9	16.4
	1	76.2	57.5	49.5	51.3	57.3
	1.5	99.8	82.9	87.9	85.1	88.5
120	0	5.1	5.3	5.4	5.2	5.2
	0.1	6.0	5.4	6.2	5.2	5.8
	0.2	8.7	7.7	7.5	6.8	6.8
	0.5	31.1	21.3	19.6	19.2	18.2
	1	90.2	74.5	63.1	66.6	64.3
	1.5	98.1	93.9	97.7	96.6	97.0
150	0	5.3	4.8	5.1	5.4	5.3
	0.1	6.3	5.8	6.5	5.6	6.0
	0.2	9.7	8.1	8.5	6.8	7.4
	0.5	43.2	22.1	22.9	21.9	21.2
	1	95.3	77.5	79.9	80.4	88.1
	1.5	100.0	97.8	98.0	99.3	99.6
180	0	5.2	5.2	4.7	5.1	5.6
	0.1	6.2	5.8	5.8	5.5	5.7
	0.2	11.2	8.8	9.5	7.7	7.9
	0.5	47.3	30.1	28.4	25.2	22.9
	1	99.0	83.2	85.5	82.5	85.6
	1.5	100.0	99.1	99.4	99.5	99.9

Table III. Type I error rate and power (%) of the proposed (LRT_F) and existing (χ^2_Z and χ^2_{CF}) approaches for the equality test of conditional correlations when the link function is misspecified ($\gamma_0 = 0.5$)

n	γ_1	Probability of Rejection				
			n=1:1		n=1:1:1	
		LRT_F	χ^2_Z	χ^2_{CF}	χ^2_Z	χ^2_{CF}
30	0	5.8	5.0	6.1	5.2	7.1
	0.1	6.8	4.6	6.6	5.0	6.8
	0.2	8.3	5.3	7.7	4.9	7.9
	0.5	41.1	8.2	9.2	7.7	8.7
	1	75.4	18.9	28.1	13.7	19.8
	1.5	98.8	49.8	37.1	27.9	41.2
60	0	5.5	5.2	5.5	5.1	5.8
	0.1	9.7	5.4	5.6	5.7	6.5
	0.2	18.2	6.3	6.9	10.1	6.8
	0.5	78.5	12.3	12.2	12.5	13.4
	1	97.6	29.6	42.7	36.9	40.9
	1.5	99.3	62.9	77.1	72.3	72.2
150	0	5.1	4.8	5.1	5.4	5.3
	0.1	14.4	5.8	6.5	5.6	6.0
	0.2	49.2	8.1	8.5	6.8	7.4
	0.5	99.5	22.1	22.9	21.9	21.2
	1	100.0	77.5	79.9	80.4	88.1
	1.5	100.0	97.8	98.0	99.3	99.6

†The true correlation is generated from a probit function, while the estimated link function in all approaches is based on Fisher z-transformation.

Table IV. Type I error rate and power (%) of the proposed (LRT_F) and existing (χ^2_Z and χ^2_{CF}) approaches for equality test of conditional correlations when data distribution is misspecified ($\gamma_0 = 0.5$)

n	γ_1	Probability of Rejection				
			n=1:1		n=1:1:1	
		LRT_F	χ^2_Z	χ^2_{CF}	χ^2_Z	χ^2_{CF}
30	0	3.9	3.1	4.4	3.4	4.8
	0.1	4.1	3.6	4.6	3.9	4.9
	0.2	5.0	6.3	5.5	3.5	4.7
	0.5	9.2	18.5	6.7	4.8	6.6
	1	23.8	34.5	22.6	7.5	20.4
	1.5	69.9	51.8	51.0	20.3	30.6
60	0	3.1	3.2	3.8	2.7	3.5
	0.1	4.1	3.3	3.7	3.0	4.1
	0.2	4.6	4.4	4.6	3.5	4.1
	0.5	12.9	10.5	12.1	7.0	9.5
	1.0	72.2	42.1	31.0	27.9	40.7
	1.5	97.3	86.5	77.2	75.8	70.6
150	0	3.5	3.3	2.8	2.7	3.1
	0.1	3.6	3.5	3.4	3.0	3.1
	0.2	7.2	5.8	5.9	4.2	3.9
	0.5	34.5	26.5	24.4	17.7	19.0
	1.0	97.7	83.8	80.9	70.0	77.8
	1.5	100.0	99.7	99.2	99.2	99.6

‡The simulated data are Pearson type II distributed, while all approaches assume bivariate normal distribution.

Table V. Parameter estimation in HOS data using the proposed approach to modeling conditional correlation between PCS and MCS

Param	Variable	Estimation	Standard Error	Degree of Freedom	t	Pr	Lower CL	Upper CL
γ_0	INTERCEPT	0.194	0.020	28715	9.84	$< .0001$	0.155	0.232
γ_1	AGE	-0.024	0.012	28715	-1.96	0.0497	-0.048	-0.001
γ_2	AGE^2	0.027	0.008	28715	3.58	0.0003	0.012	0.042
γ_3	AGE^3	-0.010	0.004	28715	-2.33	0.0199	-0.018	-0.001
γ_4	MARRIED	0.041	0.016	28715	2.61	0.0091	0.010	0.072
γ_5	GENDER	-0.124	0.017	28715	-7.27	$< .0001$	-0.158	-0.091
γ_6	RACE	0.020	0.021	28715	0.94	0.3461	-0.021	0.062