

A consistent local linear estimator of the covariate adjusted correlation coefficient

Danh V. Nguyen ¹

Division of Biostatistics, University of California

Davis, CA 95616, USA

Damla Şentürk

Department of Statistics, Pennsylvania State University

University Park, PA 16802, USA

February 14, 2009

Abstract

Consider the correlation between two random variables (X, Y) , both not directly observed. One only observes $\tilde{X} = \phi_1(U)X + \phi_2(U)$ and $\tilde{Y} = \psi_1(U)Y + \psi_2(U)$, where all four functions $\{\phi_l(\cdot), \psi_l(\cdot), l = 1, 2\}$ are *unknown/unspecified* smooth functions of an observable covariate U . We consider consistent estimation of the correlation between the unobserved variables X and Y , adjusted for the above general dual additive and multiplicative effects of U , based on the observed data $(\tilde{X}, \tilde{Y}, U)$.

Keywords: Conditional correlation; Local method of moments; Nonparametric partial correlation; Nonparametric regression; Pearson correlation.

¹Corresponding author. *Email addresses:* ucdnguyen@ucdavis.edu (D.V. Nguyen), dsenturk@stat.psu.edu (D. Şentürk)

1 Introduction

Let X and Y be random variables that are not directly observed, but are only observed after the influence of an observable third covariate U . The exact effects of U on X and Y are unknown. More precisely, the observable versions of (X, Y) are denoted (\tilde{X}, \tilde{Y}) and they are given by

$$\tilde{X} = \phi_1(U)X + \phi_2(U) \quad \text{and} \quad \tilde{Y} = \psi_1(U)Y + \psi_2(U), \quad (1)$$

where $\{\phi_l(\cdot), \psi_l(\cdot), l = 1, 2\}$ are unknown smooth functions of U . The meaning/interpretation of (1) is that there are distinct general dual additive and multiplicative effects of U on X and Y . These effects are general since the four functions can all be different and are left unspecified. The interest is to estimate the correlation between the unobserved variables X and Y , denoted ρ_{XY} , adjusted for the general effects of U in (1), based on n copies of the observable variables $(\tilde{X}, \tilde{Y}, U)$. The correlation ρ_{XY} under the dual additive and multiplicative adjustment (1) was introduced by Şentürk et al. (2008) and is called covariate adjusted correlation. Şentürk et al. (2008) showed that these general effects of U can be adjusted for, in estimating ρ_{XY} , by localization via stratification/binning of the data (\tilde{X}, \tilde{Y}) with respect to the support of U . In this paper, we consider a different localization strategy using local linear/polynomial regression. The data application which motivates model (1) involves the observed variables \tilde{X} as the observed mRNA levels of the fragile X mental retardation 1 (*FMR1*) X-linked gene and \tilde{Y} as the length of the CGG trinucleotide repeat expansion in the promotor region of the gene. For female premutation carriers (55 to 200 CGG repeats), the underlying association between CGG size and mRNA level of interest needs to be adjusted for the protective effects from one normal X chromosome. A measure that quantifies this protective effect is the activation ratio (U), which is the proportion of normal X chromosomes (Şentürk et al., 2008; Tassone et al., 2000). Furthermore, a previously proposed parametric adjustment decomposes the observed

mRNA levels into two parts, one for the carrier chromosome and one for the normal allele: $\tilde{X} = (1 - U)X + aU$, where U is activation ratio and a is the fixed mean level of mRNA expression in normal individuals (≈ 1.42 ; Tassone et al., 2000). Note that this adjustment form is a special case of the general form (1).

The estimation of ρ_{XY} of interest is facilitated by the simple, but key, observation that the (local) conditional correlation $Corr(\tilde{X}, \tilde{Y}|U = u)$ is equal to ρ_{XY} (constant) under (1) when U is independent of (X, Y) and when $\phi_1(\cdot)$ and $\psi_1(\cdot)$ are of the same sign. The later condition is satisfied when both $\phi_1(\cdot)$ and $\psi_1(\cdot)$ are positive, for instance. In such a case, the implication is that the observed measurements are positively correlated with what we want to measure, i.e. the correlation between X and \tilde{X} and between Y and \tilde{Y} are positive. Thus, since $Corr(\tilde{X}, \tilde{Y}|U = u) = \rho_{XY}$ is constant under (1), we target the conditional moment/expectation terms involved in $Corr(\tilde{X}, \tilde{Y}|U = u)$ directly. We note that the converse does not hold, i.e. positively correlated observed and unobserved variables does not imply that $\phi_1(\cdot)$ and $\psi_1(\cdot)$ are positive. In any applications, this assumption needs to be considered (justified) with the relevant subject-specific/scientific knowledge. In the aforementioned data application, for instance, both observed (\tilde{X}, \tilde{Y}) , mRNA levels and CGG length, and their activation ratio adjusted versions (X, Y) can only be positive implying positive functions $\phi_1(\cdot)$ and $\psi_1(\cdot)$.

This paper is organized as follows. The proposed covariate adjusted estimator of ρ_{XY} is described next in Section 2, where we also provide a short proof of consistency. We illustrate the computation involved with a numerical example in a simulation study in Section 3. In Section 3 we also compare the proposed method with the aforementioned binning approach of Şentürk et al. (2008) in a simulation study and illustrate the method with the *FMR1* female premutation carrier data. We conclude in Section 4 with a brief discussion of the methods and assumptions.

2 Local linear covariate-adjusted correlation estimator

Let the n copies of $(\tilde{X}, \tilde{Y}, U)$ be denoted $\{(\tilde{X}_i, \tilde{Y}_i, U_i)\}_{i=1}^n$, for n individuals. The corresponding unobserved variables (X, Y) are defined to be the parts of \tilde{X} and \tilde{Y} that are independent of U . Of interest is the correlation coefficient between X and Y , ρ_{XY} , after adjusting for the general effects of U in (1) based on the observable variables $\{\tilde{X}_i, \tilde{Y}_i, U_i\}_{i=1}^n$. Further, let $\tilde{\rho}(u)$ be the correlation between \tilde{X} and \tilde{Y} given $U = u$, defined by

$$\tilde{\rho}(u) \equiv \text{Corr}(\tilde{X}, \tilde{Y}|U = u) = \text{Cov}(\tilde{X}, \tilde{Y}|U = u) / \{\text{Var}(\tilde{X}|U = u)\text{Var}(\tilde{Y}|U = u)\}^{1/2}.$$

Note that conditioning on $U = u$ and based on the invariance of ρ_{XY} to linear transformation, it follows directly from the adjustments (1) that

$$\tilde{\rho}(u) = \rho_{XY},$$

which holds when $\phi_1(u)$ and $\psi_1(u)$ are of the same sign. Thus, within a neighborhood of u , the correlation between the observed variables \tilde{X} and \tilde{Y} , denoted $\rho_{\tilde{X}\tilde{Y}}$, will target ρ_{XY} of interest. The proposed estimator of ρ_{XY} , based on this relationship, is an average of local regression estimates of the moments in $\tilde{\rho}(u)$.

Thus, we define the following estimator of the correlation between \tilde{X} and \tilde{Y} within a neighborhood of u ,

$$r_u = \frac{\hat{\mu}_{\tilde{X}\tilde{Y}}(u) - \hat{\mu}_{\tilde{X}}(u)\hat{\mu}_{\tilde{Y}}(u)}{[\{\hat{\mu}_{\tilde{X}^2}(u) - \hat{\mu}_{\tilde{X}}^2(u)\}\{\hat{\mu}_{\tilde{Y}^2}(u) - \hat{\mu}_{\tilde{Y}}^2(u)\}]^{1/2}}, \quad (2)$$

where the individual terms in (2) are nonparametric regression estimates of $\mu_{\tilde{X}\tilde{Y}}(u) = E(\tilde{X}\tilde{Y}|U = u)$, $\mu_{\tilde{X}}(u) = E(\tilde{X}|U = u)$, $\mu_{\tilde{Y}}(u) = E(\tilde{Y}|U = u)$, $\mu_{\tilde{X}^2}(u) = E(\tilde{X}^2|U = u)$ and $\mu_{\tilde{Y}^2}(u) = E(\tilde{Y}^2|U = u)$, respectively. More precisely, the individual moments estimates in (2) can be obtained by fitting the local regression of \tilde{Y}^* on U ,

$$\tilde{Y}_i^* = \mu(u_i) + \epsilon_i,$$

where $\mu(u_i)$ is a smooth function and ϵ is a mean zero error term. For the current application, \tilde{Y}_i^* is taken to be $\tilde{X}_i\tilde{Y}_i$, \tilde{X}_i , \tilde{Y}_i , \tilde{X}_i^2 or \tilde{Y}_i^2 corresponding to the conditional expectation being estimated in (2). We fit the above local regressions by minimizing a locally weighted least squares criterion: $\sum_{i=1}^n K_h(U_i - u)[\tilde{Y}_i^* - \alpha_0 - \alpha_1(U_i - u)]^2$, where K denotes a specified kernel function with bandwidth h , $K_h(\cdot) = K(\cdot/h)/h$ and $\{\alpha_0, \alpha_1\}$ are fixed coefficients. The bandwidth h is chosen by minimizing the generalized cross-validation (Wahba, 1977; Craven and Wahba, 1979) criterion: $\text{GCV}(h) = n^{-1} \sum_{i=1}^n [\tilde{Y}_i^* - \hat{\mu}(u_i)]^2 / [1 - n^{-1} \text{tr}(H)]$, where H is the hat matrix. Extension of the least squares criterion for higher order local polynomial is straight forward.

Let r_{U_i} be the local correlation estimator (2) evaluated at $u = U_i$. Since r_{U_i} targets ρ_{XY} for all $i = 1, \dots, n$, we average these local estimates to obtain the following proposed covariate adjusted correlation estimator of ρ_{XY} ,

$$r = \frac{1}{n} \sum_{i=1}^n r_{U_i}. \quad (3)$$

The covariate adjusted estimator, r , is a consistent estimator for ρ_{XY} .

Theorem 1. *Under the technical conditions given below,*

$$r = \rho_{XY} + o_p(1).$$

Before proving the above result, we state the technical conditions and a lemma, due to Mack and Silverman (1982), that will be used. The following technical conditions are made:

(C1) The variable U is independent of X and Y , and the marginal density $f(U)$ of U has compact support, say $C(u)$, and satisfies $\inf_{u \in C(u)} f(u) > 0$, $\sup_{u \in C(u)} f(u) < \infty$.

(C2) The kernel $K(t)$ is a symmetric density function with compact support.

(C3) The functions $\{\phi_l(\cdot), \psi_l(\cdot), l = 1, 2\}$ have continuous derivatives. Furthermore, $\phi_1(\cdot)$ and $\psi_1(\cdot)$ are of the same sign.

(C4) The two observed variables of interest satisfy, $E|\tilde{X}^{2s}| < \infty$ and $E|\tilde{Y}^{2s}| < \infty$ for some $s > 2$. In addition $\sup_u \int |\tilde{x}^2|^s f(\tilde{x}, u) d\tilde{x} < \infty$ and $\sup_u \int |\tilde{y}^2|^s f(\tilde{y}, u) d\tilde{y} < \infty$, where $f(\tilde{x}, u)$ and $f(\tilde{y}, u)$ denote the joint densities of (\tilde{X}, U) and (\tilde{Y}, U) , respectively.

(C5) $h \rightarrow 0$, $nh/\log h \rightarrow \infty$, and $n^{2\epsilon-1}h \rightarrow \infty$ as $n \rightarrow \infty$, for some $\epsilon < 1 - s^{-1}$, where s is as given in Condition (C4).

The following Lemma will be used to prove Theorem 1.

Lemma 1. *Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent and identically distributed random vectors, where Y_i 's are scalar random variables. Assume further that $E|y^s| < \infty$ and $\sup_x \int |y|^s f(x, y) dy < \infty$, where f denotes the joint density of (X, Y) . Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then*

$$\sup_{x \in D} \left| n^{-1} \sum_{i=1}^n \{K_h(X_i - x)Y_i - E[K_h(X_i - x)Y_i]\} \right| = O_p(a_n),$$

provided that $n^{2\epsilon-1}h \rightarrow \infty$ for some $\epsilon < 1 - s^{-1}$, where $a_n = [nh/\log(1/h)]^{-1/2}$.

Proof of Lemma 1 follows immediately from the result obtained by Mack and Silverman (1982), as noted by Fan and Zhang (1999).

Proof of Theorem 1.

Note that in the local linear estimators of equation (2) $\hat{\mu}_{\tilde{Y}^*}(u)$ have the form $(T_4 T_3 -$

$T_5 T_2)/(T_1 T_3 - T_2^2)$, where each term T_1, \dots, T_5 satisfy

$$\begin{aligned}
T_1 &= n^{-1} \sum_{i=1}^n K_h(U_i - u) = f(u) \int K(t) dt + o_p(1), \\
T_2 &= n^{-1} \sum_{i=1}^n K_h(U_i - u)(U_i - u) = f(u) \int tK(t) dt + o_p(1), \\
T_3 &= n^{-1} \sum_{i=1}^n K_h(U_i - u)(U_i - u)^2 = f(u) \int t^2 K(t) dt + o_p(1), \\
T_4 &= n^{-1} \sum_{i=1}^n K_h(U_i - u) \tilde{Y}_i^* = \mu_{\tilde{Y}^*}(u) f(u) \int K(t) dt + o_p(1), \\
T_5 &= n^{-1} \sum_{i=1}^n K_h(U_i - u) \tilde{Y}_i^*(U_i - u) = \mu_{\tilde{Y}^*}(u) f(u) \int tK(t) dt + o_p(1),
\end{aligned}$$

uniformly in u by Lemma 1. Hence $\hat{\mu}_{\tilde{Y}^*}(u)$ targets $\mu_{\tilde{Y}^*}(u)$ uniformly in u , where \tilde{Y}_i^* can be taken to be $\tilde{X}_i \tilde{Y}_i$, \tilde{X}_i , \tilde{Y}_i , \tilde{X}_i^2 or \tilde{Y}_i^2 . Thus, the following holds uniformly in u for r_u , given in (2):

$$\begin{aligned}
r_u &= \frac{\mu_{\tilde{X}\tilde{Y}}(u) - \mu_{\tilde{X}}(u)\mu_{\tilde{Y}}(u)}{\{\sigma_{\tilde{X}}^2(u)\sigma_{\tilde{Y}}^2(u)\}^{1/2}} + o_p(1) = \text{Corr}(\tilde{X}, \tilde{Y} | U = u) + o_p(1) \\
&= \rho_{XY} + o_p(1),
\end{aligned}$$

and Theorem 1 follows.

3 Simulation studies and data example

3.1 Simulation results

To illustrate the computation associated with the covariate-adjusted correlation, consider $(X, Y)^T$ to be bivariate normal with mean $\mu = (2, 3)^T$, $\text{Var}(X) = 3$, $\text{Var}(Y) = 4$ and the correlation of interest is $\rho_{XY} = 0.25$. The observed data (\tilde{X}, \tilde{Y}) is obtained as follows. We have $\tilde{X} = \phi_1(U)X + \phi_2(U)$ with $\phi_1(U) = 5 \log(U + 1)$, $\phi_2(U) = 3U$ and $U \sim \text{Uniform}[2, 5]$. Similarly, $\tilde{Y} = \psi(U)Y + \psi(U)$ with $\psi(U) = \exp(U)/U^2$ and $\psi(U) = -2U^2$. The observed correlation is $\rho_{\tilde{X}\tilde{Y}} \approx -0.14$. Figure 1 displays $n = 200$ pairs of $\{\tilde{X}_i, \tilde{Y}_i\}_{i=1}^n$ as well as the

unobserved data $\{X_i, Y_i\}_{i=1}^n$. The local correlation (2) is obtained as described in Section 2 above.

Figure 2 displays the covariate adjusted correlation estimates for 200 Monte Carlo data sets for sample sizes $n = 50, 75, 100$ and 200 . The mean (standard deviation) over the simulation corresponding to sample sizes 50 to 200 are: 0.276 (0.271), 0.260 (0.147), 0.255 (0.133) and 0.256 (0.078), respectively. The estimates target the true correlation of $\rho = 0.25$ with decreasing standard deviation as n increases, as expected.

Next, we compare the proposed local linear estimator of the covariate adjusted correlation to another approach based on binning the data (Şentürket al. 2008). Briefly, the binning method is as follows. The observed data is binned with respect to U . That is, the range of U is divided into m equidistant intervals, referred to as bins and denoted by B_1, \dots, B_m . Let L_j denote the number of subjects falling into bin j , $1 \leq j \leq m$. All the data points falling into bin j are denoted $(U'_{jk}, \tilde{X}'_{jk}, \tilde{Y}'_{jk})$, for subjects $1 \leq k \leq L_j$. Observations within any bin are marked by a prime. The Pearson correlation is then calculated using data in each bin. The estimate of the covariates adjusted correlation is then obtained by averaging the correlation coefficients from the m bins. More formally, the correlation between \tilde{X} and \tilde{Y} within bin j is,

$$r_{j,\text{bin}} = \frac{M_{\tilde{X}\tilde{Y},j} - M_{\tilde{X},j}M_{\tilde{Y},j}}{\sqrt{M_{\tilde{X}^2,j} - M_{\tilde{X},j}^2}\sqrt{M_{\tilde{Y}^2,j} - M_{\tilde{Y},j}^2}},$$

where $M_{\tilde{X}\tilde{Y},j} = L_j^{-1} \sum_{k=1}^{L_j} \tilde{X}'_{jk} \tilde{Y}'_{jk}$, $M_{\tilde{X},j} = L_j^{-1} \sum_{k=1}^{L_j} \tilde{X}'_{jk}$, $M_{\tilde{Y},j} = L_j^{-1} \sum_{k=1}^{L_j} \tilde{Y}'_{jk}$, $M_{\tilde{X}^2,j} = L_j^{-1} \sum_{k=1}^{L_j} \tilde{X}'_{jk}{}^2$ and $M_{\tilde{Y}^2,j} = L_j^{-1} \sum_{k=1}^{L_j} \tilde{Y}'_{jk}{}^2$. The estimate of the covariate adjusted correlation is given by the following weighted average

$$r_{\text{bin}} = \sum_{j=1}^m \frac{L_j}{n} r_{j,\text{bin}}.$$

The bin-specific correlation estimates are weighted ($\{L_j/n\}$) relative to the number of observations in each bin.

Figure 3 (top) displays the relative bias of r (local linear) and r_{bin} , where the true correlation (as before) is 0.25. Although the observed maximum absolute biases of both methods are small (less than ~ 0.024), the bias for the local linear estimate is relatively smaller (e.g. at $n = 75$ and 100). The bias decline with increasing n , as expected, and becomes similar for $n = 200$ in the current simulation. The bottom plot of Figure 3 displays the mean square error (MSE) of the two estimators. The variances of the two estimators (not displayed) closely track the given MSE. Thus, the covariate adjusted estimator based on data binning appears to have smaller MSE (and variance).

We note here, as previously pointed out in Şentürk et al. (2008), that the binning approach requires the specification of the number of bins m . Aside from the basic guideline that there should be sufficient data within each bin to compute the correlation coefficient, there is no automatic way to select this parameter and Şentürk et al. (2008) recommended that in applications a sensitivity analysis be used for examining the variability in the estimates for different choices of m . Although for the simulation studies we selected m based on MSE, this approach is not feasible in practice (since the MSE would not be known). Thus, one advantage of the proposed local linear estimator is that the bandwidth parameter choice is chosen by generalized cross-validation.

3.2 Data example

To illustrate the proposed method, we consider data consisting of molecular measurements, $\widetilde{\text{CGG}}$, $\widetilde{\text{mRNA}}$ and $U =$ activation ratio, for $n = 165$ female premutation carriers (Şentürk et al., 2008). The aim is to estimate the correlation $\rho_{XY} \equiv \rho_{\text{mRNA,CGG}}$, the activation ratio-adjusted correlation between the mRNA level and CGG repeat size. The observed variables, $\widetilde{\text{CGG}}$, $\widetilde{\text{mRNA}}$ and activation ratio, range between (57, 138) repeats, (0.78, 6.3) and (0.19, 0.91), respectively. As previously reported in Şentürk et al. (2008), the correlation estimate based on data binning is $r_{\text{bin}} = 0.37$ and this estimate is quite similar for the number of bins ranging

in $m = 17$ and $m = 25$ (corresponding to $\sim 7 - 10$ observations per bin). Estimation of $\rho_{XY} \equiv \rho_{\text{mRNA,CGG}}$ based on the local linear approach provide a similar estimate of $r = 0.39$.

As discussed in Tassone et al. (2000), without adjusting for the protective effects of the normal chromosome in female premutation carriers (i.e. ignoring activation ratio), the strength of association between CGG size and mRNA level is weaker (unadjusted correlation ~ 0.29), roughly half that seen in male premutation carriers (where activation level is not an issue). The activation ratio-adjusted correlation estimate is higher and more similar to the level of association seen in male premutation carriers.

4 Conclusion

In this work, we considered a local linear (polynomial) estimator of the covariate adjusted correlation between two unobserved variables X and Y , adjusted for the general (unknown) effects of a third observable covariate. We showed that the proposed estimator is consistent. In calculating the estimator, the bandwidth can be chosen by minimizing the generalized cross-validation criterion. The implementation is easy and can be based on available routine for nonparametric regression. A simple choice is local regression and we implemented the proposed method using the publicly available R package `locfit` (<http://www.locfit.info/>). The current estimation method has the advantage that the bandwidth choice can be chosen more systematically using GCV and the proposed method can augment the estimation approach based on binning the data previously proposed.

Acknowledgment

Support for this work includes the National Institute of Health (NIH) grants UL1RR024922, RL1AG032119 and RL1AG032115, National Institute of Child Health and Human Development grant HD036071, and grant UL1 RR024146 from the National Center for Research Re-

sources (NCCR), a component of NIH.

References

- Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31, 377-403.
- Fan, J. and Zhang, W., 1999. Statistical estimation in varying coefficient models. *Annals of Statistics* 27, 1491–1518.
- Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 61, 405–415.
- Şentürk, D., Nguyen, D. V., Tassone, F., Hagerman, R. J., Carroll, R. J., Hagerman, P. J., 2008. Covariate adjusted correlation analysis with application to *FMR1* premutation female carrier Data, *Biometrics*, in-press.
- Tassone, F., Hagerman, R. J., Chamberlain, W. D. and Hagerman, P. J. 2000. Transcription of the *FMR1* gene in individuals with fragile X syndrome. *American Journal of Medical Genetics* 97, 195–203.
- Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In Krishnaiah, P. R. (Ed.), *Applications of Statistics*. North Holland, Amsterdam, p. 507-523.

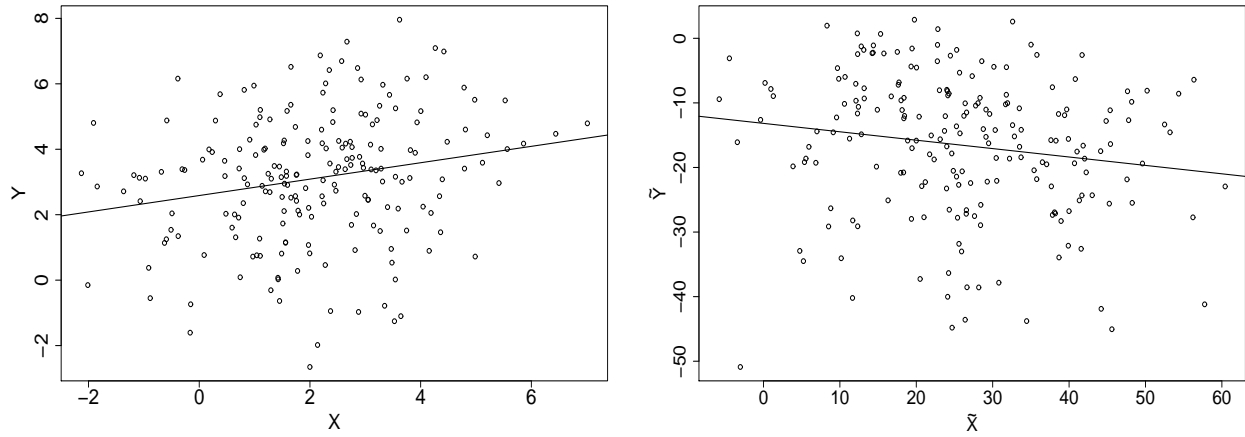


Figure 1: Data pairs (left) (X_i, Y_i) and (right) $(\tilde{X}_i, \tilde{Y}_i)$, $i = 1, \dots, n = 200$.

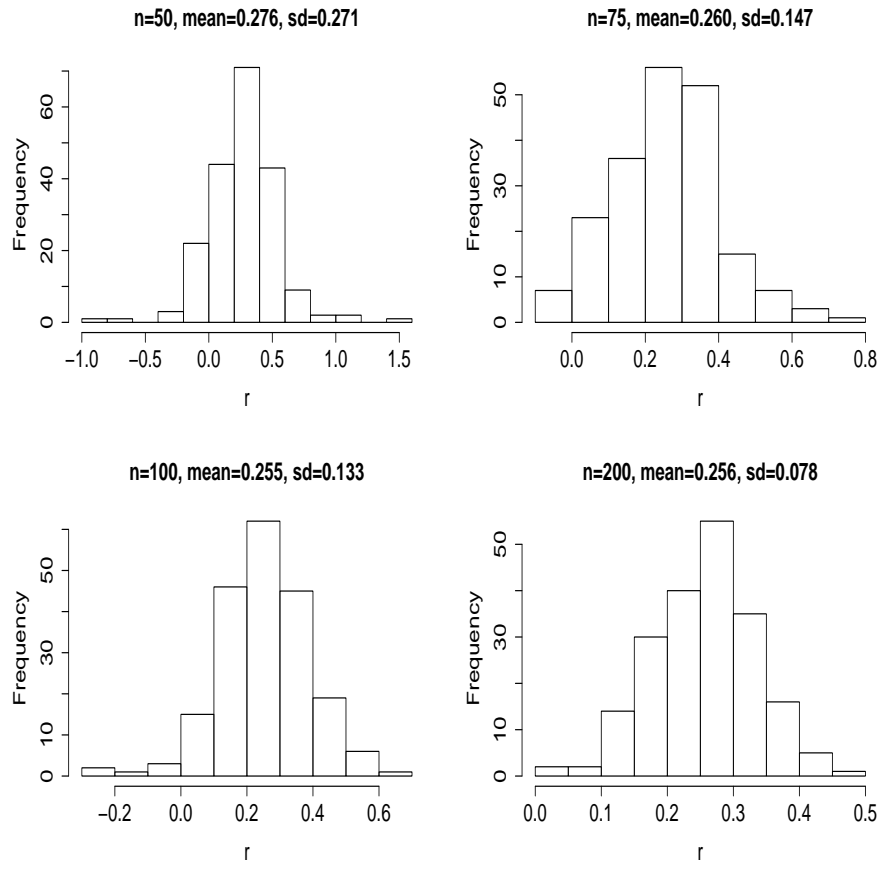


Figure 2: Covariate-adjusted correlation estimates over 200 Monte Carlo data sets of size $n = 50, 75, 100$ and 200 .

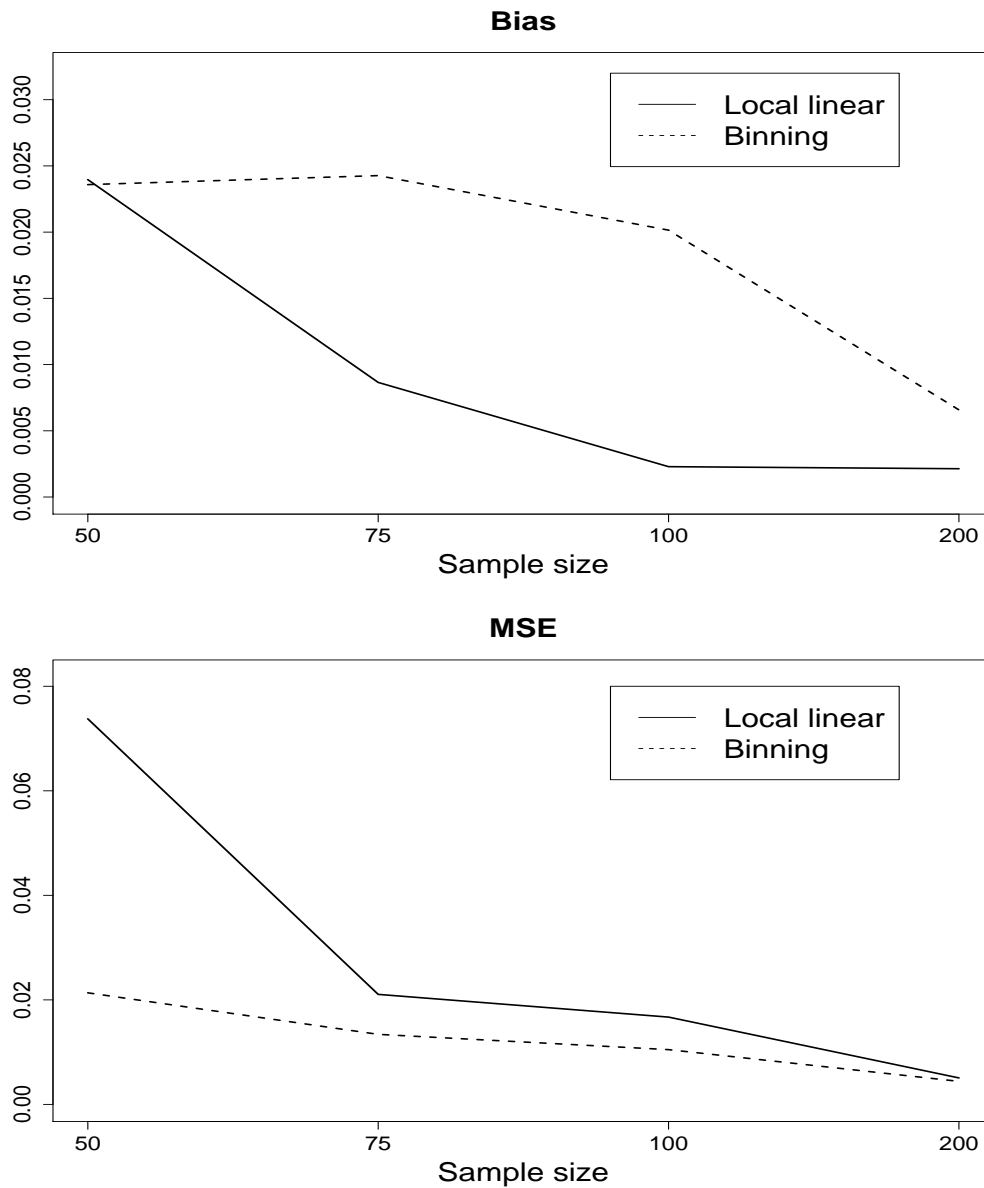


Figure 3: Bias and mean square error for the proposed local linear approach to estimate the covariate-adjusted correlation compared to the binning approach. Plotted values are averaged over 200 Monte Carlo data sets of size $n = 50, 75, 100$ and 200 .