Published in final edited form as:

Genet Epidemiol. 2012 December ; 36(8): 890–894. doi:10.1002/gepi.21671.

Loss of Power in Two-Stage Residual-Outcome Regression Analysis in Genetic Association Studies

Ronglin Che, Alison A. Motsinger-Reif*, and Chad C. Brown

Bioinformatics Research Center, Department of Statistics, North Carolina State University, Raleigh, North Carolina

Keywords

loss of power; two-stage regression; degrees of freedom

In the area of genetic epidemiology, investigating the association between genetic risk factors and phenotypic outcomes is an important challenge. Often in such studies, the analysis approach needs to control for potentially confounding factors, such as clinical, demographic, or environmental variables. Multiple linear regression (MLR) is often used for such analysis when the outcome of interest is a quantitative trait. Under this model, the effects of genetic factors are estimated effectively and covariates are adjusted correctly. In large-scale genome-wide studies with a relatively large number of genetic and environmental factors present, some alternative analytic strategies are being used to address the issues of covariate adjustment. One such approach is a two-stage residual-outcome regression analysis (2SR). At stage one, the outcome is regressed on all the covariates. At stage two, the residual-adjusted outcome is then regressed on genetic factors. 2SR has been employed in genetic association studies [Choy et al., 2008], as well as genetic linkage studies [Slager and Iturria, 2003]. Zeegers and colleagues have demonstrated that the 2SR has equal power with other covariate-adjusted methods in genetic linkage studies [Zeegers et al., 2004], but the use of such a two-stage approach in association analyses is less well understood. While there may be some potential advantages with the 2SR approach in both ease of use and computational and data management efficiency, some issues, such as the bias of estimates, power, and type I error, need to be thoroughly investigated.

Recently, an article by Demissie and Cupples provided an insightful and important comparison of the 2SR and MLR analysis in the context of genetic association studies [Demissie and Cupples, 2011]. The authors discovered that when a single nucleotide polymorphism (SNP) and a covariate are correlated, the 2SR method results in a biased estimate of genotypic effect. Furthermore, this bias is always toward the null hypothesis and increases as the squared correlation between SNP and covariate increases. This was an

important result, and in the current manuscript we discuss their results, introduce a few points of clarification, and perform a simulation study to demonstrate how the comparison between the two methods might extend to more realistic data.

Here the MLR model with two variables S (SNP) and C (covariate), and outcome Y is given by

$$\begin{aligned} Y &= \beta_0 + \beta_1 S + \beta_2 C + \varepsilon \\ \beta_1 &= \rho_{Y|C} \frac{\sigma_{Y|C}}{\sigma_{S|C}}, \end{aligned} \quad (1)$$

where the parameter $\rho_{Y|C}$ is the population partial correlation between Y and S adjusting for C ; $\sigma_{Y|C}$ and $\sigma_{S|C}$ are the conditional standard deviations (SDs) of Y and S given C , respectively [Kleinbaum, 2007]. The least squares estimate of β_1 is equivalent to substituting

sample estimates for the parameters in 1, giving $\hat{\beta}_1 = r_{Y|C} \frac{S_{Y|C}}{S_{S|C}}$. To test $H_0: \beta_1 = 0$, the test

statistic $\hat{F} = \frac{r_{Y|C}^2}{(1 - r_{Y|C}^2)/(n-3)}$ is compared with an F distribution having $df(1, n-3)$. The

“true value” of F using model parameters is $F = \frac{\rho_{Y|C}^2}{(1 - \rho_{Y|C}^2)/(n-3)}$. In a 2SR model,

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 C + e \\ Z &= Y - \hat{Y} = \beta_0^* + \beta_1^* S + \varepsilon^* \\ \beta_1^* &= \rho_{(Y|C)S} \frac{\sigma_{Y|C}}{\sigma_S} = \rho_{ZS} \frac{\sigma_Z}{\sigma_S}, \end{aligned}$$

where $\rho_{(Y|C)S}$ is the population semipartial correlation between Y and S with only Y adjusting for C ; σ_S is the SD of S . The corresponding estimate of β_1^* is

$\hat{\beta}_1^* = r_{(Y|C)S} \frac{S_{Y|C}}{S_S} = r_{ZS} \frac{S_S}{S_S}$. To test $H_0: \beta_1^* = 0$, we compare $\hat{F}^* = \frac{r_{(Y|C)S}^2}{(1 - r_{(Y|C)S}^2)/(n-2)}$, with an F distribution having $df(1, n-2)$. The corresponding population parameter is

$$F^* = \frac{\rho_{(Y|C)S}^2}{(1 - \rho_{(Y|C)S}^2)/(n-2)}.$$

Demissie and Cupples incorrectly labeled the true values F and F^* as test statistics, and gave the wrong degrees of freedom for F as $(1, n-2)$, producing the equality:

$$\frac{F^*}{F} = 1 - \frac{\rho_{SC}^2}{1 + \rho_{SC}^2 \rho_{Y|C}^2 - \rho_{Y|C}^2}.$$

In actuality, this equality only approximately holds for the test statistics $\hat{F}$ and $\hat{F}^*$, as the appropriate correlation parameters need to be estimated using data. Using the correct degrees of freedom, we produce the approximate equality:

$$d \approx \frac{\hat{F}^*}{\hat{F}} = \frac{n-2}{n-3} \left(1 - \frac{r_{SC}^2}{1 + r_{SC}^2 r_{S|C}^2 - r_{S|C}^2} \right) = \frac{n-2}{n-3} \left(1 - \frac{r_{SC}^2}{\hat{m}} \right). \quad (2)$$

Although the multiplier $(n-2)/(n-3)$ should be close to 1 for moderate sample sizes, this ratio only applies to a single environmental covariate. In general, it should be noted that for testing the genetic effect (1 SNP) and k covariates in a MLR model, the F -test statistic has an F distribution with $df(1, n-k-2)$. However, the F^* statistic still has the F distribution with $df(1, n-2)$. This implies the multiplier d could be generalized to

$d \approx \frac{n-2}{n-k-2} \left(1 - \frac{R_{SC}^2}{1 + R_{SC}^2 r_{S|C}^2 - r_{S|C}^2} \right) = \frac{n-2}{n-k-2} \left(1 - \frac{R_{SC}^2}{\hat{m}} \right)$, where R_{SC}^2 is R^2 of fitting S on C and C could be more than one dimension. If C is only one dimension, then R_{SC}^2 equals the squared correlation between S and C , that is r_{SC}^2 . Based on this relationship, the multiplier d depends not only on R_{SC}^2 and $r_{S|C}^2$, but also on the sample size and the number of covariates.

The accuracy of this approximation was tested using simulation, where S and C were generated as multivariate normal, and Y was generated based on C only. Two choices of sample size (100 and 500), two choices of the number of covariates (1 and 20), and three choices of association between S and C (negligible, moderate, and strong) led to 12 combinations in total. Sample size of 1,000 was also generated and results are the same. Details were not shown. For each combination, 1,000 data sets were generated with $\hat{F}$ and $\hat{F}^*$ calculated for each set. These statistics are plotted in Figure 1 with the diagonal line indicating $\hat{F}^* = d \hat{F}$. The mean of R_{SC}^2 for each row was calculated. Visually, from the spread of points, we observe that for strong associations between S and C , or for small sample sizes, the approximation in 2 is poor. However, for moderate samples sizes and modest associations likely to be encountered in practice, the approximation is good. Furthermore, when the number of covariates is relatively large (20) compared to the sample size (100), d (1.256) could be greater than 1, and the calculated R_{SC}^2 (statistic) should also be greater than the theoretic version of R_{SC}^2 (parameter). To clarify their difference, “true R_{SC}^2 ” is defined as the proportion of variation of S that could be explained by all the covariates in our simulation design, that is one minus the conditional variance of S given C divided by the unconditional variance of S . Since d is calculated on the “true R_{SC}^2 ” which is smaller than R_{SC}^2 (statistic), the approximation is off-diagonal for Figure 1 panel c with $k = 20$ and $n = 100$. In addition, since Y is independent of S in the simulation design, we could also investigate the type I error rate for each simulated combination. Type I error rate was defined as the proportion of rejecting the null hypothesis (P -values less than 0.05) across 1,000 replicates. Results show that the type I error rates for MLR is approximately 0.05 across all combinations, as expected. However, the averaged type I error rates for 2SR decrease sharply from 0.044, 0.022 to 0, when the associations between S and C increase

from negligible, moderate to strong. It is clearly that 2SR may be too conservative when there is an association between S and C .

As a minor note, we also want to make a clarification about the semipartial correlation $r_{(Y|C)S}$. The $r_{(Y|C)S}$ is a correlation between Y and S with only Y adjusted for C and $r_{(Y|C)S} =$

$r_{Y-\hat{Y},S} = r_{Z,S}$. In the original paper, the authors suggest that $\rho_{(Y|C)S} = \rho_{ZS} = \frac{\rho_{YS} - \rho_{YC}\rho_{SC}}{\sqrt{1 - \rho_{YC}^2}}$. However, a more appropriate version should be

$$\rho_{(Y|C)S} = \rho_{ZS} = \frac{\rho_{YS} - \rho_{YC}\rho_{SC}}{\sqrt{1 - \rho_{YC}^2}} = \rho_{YS|C} \sqrt{1 - \rho_{SC}^2}.$$

In addition to the concerns with bias, there are other concerns with the 2SR approach. The 2SR approach estimates a different parameter than MLR, so an interpretation of effect size should take this into account. With the 2SR approach, the interpretation about the effect size of β_1^* could only be given for Z and S , rather than for Y and S , which is usually the primary interest of such an analysis. In practice, investigators should be cautious in interpreting the effect size from 2SR.

To demonstrate these additional considerations for the power comparison between MLR and 2SR, we performed a simulation study. The simulation was conducted to demonstrate the relationship among F , $\hat{F}^*$, the multiplier d , and the noncentrality parameter (effect size) of the F distribution. Based on the results of Figure 1, we observe that the approximation of 2 is good when the sample size is relatively large. Since the good performance of approximation is a prerequisite in this simulation, a sample size of 500 was chosen. The number of covariates k was set to 20, since we try to illustrate the scenario with d larger than 1. The primary goal here is to show how could d change the relative power between MLR and 2SR. The data generation details follow. First, we generated 100,000 random F deviates with df 1 and $n-k-2$. Based on the approximate relationship between the F and $\hat{F}^*$, we calculated F^* for several values of d , ranging from 0.25, 0.5, 0.75, 1, and $d_{\max}$, where $d_{\max}$ (assuming no association between S and C) equals $(n-2)/(n-k-2)$. Then we calculated the proportions that are above the critical values that would actually be used in hypothesis testing, denoted as power. The power for MLR was calculated as the proportion of F statistics that are bigger than critical values for $F(1, n-k-2)$ (a correct assumption under the null), with type I error 0.05, while the power for 2SR would be the proportion that $\hat{F}^*$ ($\approx d F$) are above the critical values for $F^*(1, n-2)$ (an incorrect assumption). The noncentrality parameter of F distribution was set as the following range: 0, 2, 4, 6, 8, 10, 12, 14, 16, where 0 represents the null hypothesis. The power comparison is shown in Figure 2. We could observe that when d is larger than 1, the 2SR is more powerful than MLR at the cost of high type I error rate (0.05381), and thus the tests are liberal. When d is equal to 1, power (Figure 2) and type I error rate for MLR (0.04918) and 2SR (0.04921) are similar. When d is smaller than 1, tests are not powerful and too conservative (low type I error rate). The type I error rates were 0.02313, 0.00533, and 0.00013, when d were 0.75, 0.5, and 0.25, respectively.

Based on the relationship $d \approx \frac{n-2}{n-k-2} (1 - \frac{R_{SC}^2}{\hat{m}})$, we should also note that as the sample size n decreases or the number of covariates k increases, the difference in the denominator

degrees of freedom between MLR ($n-k-2$) and 2SR ($n-2$) increases. In this case, d could be much greater than 1 resulting in an inflated type I error. However, for the large sample sizes and reasonable number of covariates commonly encountered in practice, d is expected to be less than 1 especially when SNP and covariate(s) are associated, leading to a loss of power for 2SR. Also, when sample sizes are large and there is no association between SNP and covariate(s), d approximates to 1 and roughly equating F and F^* , giving both methods similar power.

In light of the approximate nature of 2, and the recognition that many analyses involve more than a single environmental covariate, the relative performances of 2SR and MLR were compared in a more realistic setting using simulation. For this, data were generated using a number of continuous and discrete environmental covariates, a continuous response and a single common SNP under Hardy-Weinberg equilibrium (HWE) with a minor allele frequency (MAF) of 0.4. Three covariates $\mathbf{C} = (C_1, C_2, C_3)^T$ were associated with both SNP and potential response, that is, confounders, and three covariates $\mathbf{V} = (V_1, V_2, V_3)^T$ were associated with potential response but not the SNP values. Each SNP was generated as multinomial to approximate HWE, and the value S was coded as 0, 1, and 2, corresponding to the number of minor alleles. $(V_1, V_2)^T$ was generated as multivariate normal $N(\boldsymbol{\mu}_V, \Sigma)$,

where $\boldsymbol{\mu}_V = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, and $V_3 \sim \text{Bernoulli}(0.4)$. For confounders, we consider three settings (negligible, moderate, and strong) to represent varying degrees of association between the SNP and the confounders. $(C_1, C_2)^T$ was generated as multivariate normal $N(\boldsymbol{\mu}_{S,asso}, \Sigma)$, and $C_3 \sim \text{Bernoulli}(p_{S,asso})$, where $\boldsymbol{\mu}_{S,asso}$ and $p_{S,asso}$ were specified in Table I, and Σ was the same as previous variance. Based on the specifications of $(\mathbf{C}, \mathbf{V})$, the response Y was generated according to

$$Y = \gamma_0 + \gamma_1 C_1 + \gamma_2 C_2 + \gamma_3 C_3 + \xi_1 V_1 + \xi_2 V_2 + \xi_3 V_3 + \beta S + \varepsilon, \varepsilon \sim N(0, 1),$$

where $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \xi_1, \xi_2, \xi_3) = (0, -0.5, 0.5, 0.5, -0.5, 0.5, 0.5)$ and β (effect size) represents the different degrees of association between response and SNP, ranging from 0, 0.2, 0.4, 0.6, 0.8, and 1. For each combination, 1,000 Monte Carlo data sets were generated for sample size ranging from 50, 100, and 500. The power was defined as the proportion of data sets with statistically significant effect of SNP (P -values less than 0.05), using MLR and 2SR method, respectively. Figure 3 shows that the test of MLR is at least as powerful as that of 2SR, as expected. If there is no association between SNP and confounders, the power of MLR and 2SR are equal. As the association between SNP and confounders increases, the power of MLR is much larger than that of 2SR.

In summary, we emphasize the key conclusions of the Demissie and Cupples [2011] study, and clarify the theoretical relationship between the 2SR and MLR in the genetic association study. We expand on these points using some simple simulation examples. These results suggest that we should be cautious to apply the two-stage residual-outcome regression, since it may introduce the bias and loss of power with the presence of confounder(s).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by T32GM081057 from the National Institute of General Medical Sciences and the National Institute of Health and NCI R01 CA161608.

References

- Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, De Jager PL, Shaw SY, Wolfish CS, Slavik JM, Cotsapas C, Rivas M, Dermitzakis ET, Cahir-McFarland E, Kieff E, Hafler D, Daly MJ, Altshuler D. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.* 2008; 4(11):e1000287. [PubMed: 19043577]
- Demissie S, Cupples LA. Bias due to two-stage residual-outcome regression analysis in genetic association studies. *Genet Epidemiol.* 2011; 35(7):592–596. [PubMed: 21769934]
- Kleinbaum, DG. *Applied Regression Analysis and Other Multivariable Methods.* Australia; Belmont, CA: Brooks/Cole; 2007.
- Slager SL, Iturria SJ. Genome-wide linkage analysis of systolic blood pressure: a comparison of two approaches to phenotype definition. *BMC Genet.* 2003; 4(Suppl 1):S13. [PubMed: 14975081]
- Zeegers M, Rijdsdijk F, Sham P. Adjusting for covariates in variance components QTL linkage analysis. *Behav Genet.* 2004; 34(2):127–133. [PubMed: 14755177]

Fig. 1.

A comparison of test statistics ($\hat{F}$ and $\hat{F}^*$) and parameters (F and F^*) in multiple linear regression (MLR) and two-stage residual-outcome regression (2SR) across 1,000 replicates.

(R^2_{SC} is the mean of R^2_{SC} for each row, across all replicates.)

Fig. 2.

A comparison of multiple linear regression (MLR) and two-stage residual-outcome regression (2SR) analysis in terms of power across 100,000 replicates with 20 covariates and sample size of 500.

Fig. 3. Power of multiple linear regression (MLR) and two-stage residual-outcome regression (2SR) analysis in a simulated example with 1 SNP + 3 confounders + 3 covariates. (R^2_{SC} is the mean of R^2_{SC} for each row, across all replicates.)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE I

A specification of $\mu_{S,asso}$ and $p_{S,asso}$ in the simulation example

Association	$p_{S,asso}$				$\mu_{S,asso}$			
	$S = 2$	$S = 1$	$S = 0$	S	$S = 2$	$S = 1$	$S = 0$	S
Negligible	0.4	0.4	0.4	$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$	
Moderate	0.6	0.4	0.2	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0.5 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		
Strong	0.8	0.4	0.1	$\begin{pmatrix} 2 \\ 2 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$		