Bias due to 2-stage residual-outcome regression analysis in genetic association studies

Serkalem Demissie[#,1] and **L. Adrienne Cupples, PhD**[1]

[1]Boston University School of Public Health, Department of Biostatistics, Boston, MA

Abstract

Association studies of risk factors and complex diseases require careful assessment of potential confounding factors. Two-stage regression analysis, sometimes referred to as residual- or adjusted-outcome analysis, has been increasingly used in association studies of single nucleotide polymorphisms (SNPs) and quantitative traits. In this analysis, first, a residual-outcome is calculated from a regression of the outcome variable on covariates and then the relationship between the adjusted-outcome and the SNP is evaluated by a simple linear regression of the adjusted-outcome on the SNP. In this paper, we examine the performance of this 2-stage analysis as compared with multiple linear regression (MLR) analysis. Our findings show that when a SNP and a covariate are correlated, the 2-stage approach results in biased genotypic effect and loss of power. Bias is always toward the null and increases with the squared-correlation between the SNP and the covariate (ρ_{sc}^2). For example, for ρ_{sc}^2=0.0, 0.1 and 0.5, 2-stage analysis results in, respectively, 0%, 10% and 50% attenuation in the SNP effect. As expected, MLR was always unbiased. Since individual SNPs often show little or no correlation with covariates, a 2-stage analysis is expected to perform as well as MLR in many genetic studies; however, it produces considerably different results from MLR and may lead to incorrect conclusions when independent variables are highly correlated. While a useful alternative to MLR under ρ_{sc}^2=0.0, the 2-stage approach has serious limitations. Its use as *a simple substitute for MLR* should be avoided.

1. BACKGROUND

Confounding may occur when independent variables (predictors) are associated with one another and with the outcome of interest [Hennekens, et al., 1987]. If not appropriately accounted for, confounding factors can lead to biased results and misleading conclusions. Bias due to confounding can be minimized or controlled by a study design or by employing appropriate data analysis methods such as multiple regression, propensity score, or stratification analyses [Rothman & Greenland, 1998].

For an association study of a quantitative outcome and a genetic risk factor, multiple linear regression (MLR) analysis can be employed to effectively adjust for covariates to reduce variation or to minimize confounding effects. Environmental or biological variables, such as dietary factors, smoking, age, and sex, are often used as covariates in MLR with the goal of

[#]**Corresponding author:** Serkalem Demissie, PhD, Department of Biostatistics, Boston University School of Public Health BUMC, 801 Massachusetts Avenue, 315, Boston, MA 02118, Tel: 617-638-5249, Fax: 617-638-6484, demissie@bu.edu.

reducing noise in the outcome variation and estimating the genotype effect more precisely [Christenfeld et al., 2004]. As an alternative to multiple linear regression analysis, increasing numbers of studies use a 2-stage approach, where at stage one a residual variable (also referred as an 'adjusted' outcome) is obtained from a linear regression of the outcome variable on potential confounding factors and at stage two the relationship between the adjusted-outcome and the exposure variable is evaluated using a simple linear regression analysis. This approach has many attractive practical advantages over multivariable analysis (e.g., computational and data management efficiencies in large-scale genome-wide association studies). But its validity relies on the assumption that covariates and the risk factor of interest are uncorrelated. When this assumption is not met, as may be the case when a covariate under consideration is a principal components (PC) of ancestry to account for potential confounding effects of population stratification [Price et al., 2006] or another SNP in linkage disequilibrium with the primary SNP of interest (e.g., in conditional analysis), the 2-stage analysis can produce biased results because it does not fully take into account the multivariable relationship among all study-variables.

The present study was conducted to examine a 2-stage approach as compared with MLR analysis in the context of a population-based association study of a genetic marker and a quantitative trait.

2. METHODS

In this section we present a brief review of the linear regression method within the context of partial and semi-partial correlations [Kleinbaum et al., 1988] and illustrate the relationship between MLR and 2-stage analyses using two independent variables, S and C, and an outcome variable Y; where S is the exposure variable of interest (SNP genotype) and C is a potential confounding factor or covariate (genetic or environmental factor) that may or may not be associated with Y and S. Although only two independent variables are considered for illustrative purposes, the results and discussions can be generalized to studies of multiple independent variables.

2.1 Multiple Linear Regression (MLR) (a single-stage procedure)

A regression model with two independent variables, S and C, and an outcome variable Y is given by

$$Y = \beta_0 + \beta_1 S + \beta_2 C + \varepsilon \tag{1}$$

where β_i ($i = 0,1,2$) are unknown parameters that may be estimated using the least squares method which minimizes the sum of squared deviations of the residuals; and ε is a random error that accounts for unexplained random variation in Y. Following regression and correlation theories, β_i's can be represented in terms of means, standard deviations (SDs) and partial correlation coefficients [Kleinbaum et al., 1988]. Accordingly, the additive effect of the SNP (S) in Eq. (1) can be given by

$$\beta_1 = \rho_{YS|C} \frac{\sigma_{Y|C}}{\sigma_{S|C}} \tag{2}$$

Where $\rho_{YS|C}$ is the partial-correlation for Y and S adjusting for C; and $\sigma_{Y|C}$ and $\sigma_{S|C}$ are conditional standard deviations of Y and S given C, respectively. An important observation from Eq. (2) is that the parameter of effect for S is a function of the _partial_-correlation

($\rho_{YS|C}$, correlation that removes the covariate effect from *both* Y and S) and the SDs of both Y and S conditional on C.

2.2 Two-stage residual-outcome regression

In addition to the above three variables (Y, S and C), consider a residual-outcome variable (or 'adjusted' outcome) denoted by Z. In a 2-stage analysis, the residual variable ($Z = Y - \hat{Y}$) is obtained *at stage one* from a linear regression of the outcome variable on the potential confounding factor ($Y = a_0 + a_1 C + e$) and the relationship between the residual and the exposure variable S is evaluated *at stage two* using a simple linear regression model $Z = \beta_0^* + \beta_1^* S + \varepsilon^*$; where a_0 and a_1 are unknown parameters and e is a random error in the first stage and β_0^* and β_1^* are unknown parameters and ε^* is a random error in the second stage. Similar to the representation given in Eq. (2), β_1^* can be expressed as follows:

$$\beta_1^* = \rho_{ZS} \frac{\sigma_Z}{\sigma_S}$$

(3)

Note that ρ_{ZS} is the semi-partial correlation between Y and S where *only* Y is adjusted for C ($\rho_{(Y|C)S}$); σ_Z is the conditional SD of Y given C ($\sigma_{Y|C}$); and σ_S is the unconditional SD of S. Thus, the regression parameter for S in this two-stage residual-outcome analysis, Eq. (3), is a function of the <u>*semi-partial* correlation</u> (correlation that removes the covariate effect *only* from the outcome), the *conditional* SD of Y given C, and the *unconditional-SD* of S.

2.3 Relationship between 2-stage regression and MLR: Parameter Estimate

By solving Eqs. (2) and (3) it can be shown that the relationship between the regression parameters from MLR (β_1) and 2-stage (β_1^*) regression models can take the form $\beta_1^* = \beta_1(1 - \rho_{SC}^2)$, where ρ_{SC}^2 the squared correlation between S and C (see Appendix for details). Consequently, the following conclusions can be drawn:

1. 2-stage analysis is unbiased and has an identical solution to that of MLR when S is *not* associated with C:

$$\beta_1^* = \beta_1 \quad if \, \rho_{SC}^2 = 0$$

2. 2-stage analysis is biased and *underestimates* the exposure effect when S is correlated with C:

$$|\beta_1^*| < |\beta_1| \quad if \, \rho_{SC}^2 > 0$$

3. For $\beta_1 \neq 0$, the *expected* bias due to 2-stage analysis ($\beta_1 - \widehat{\beta_1^*} = \beta_1 \rho_{SC}^2$, where $\widehat{\beta_1^*}$ is the 2-stage estimator of the parameter β_1) is independent of the association between the outcome and the covariate. Thus, even when confounding is not present in the data, bias can be introduced simply as a direct consequence of performing a 2-stage analysis.

2.4 Relationship between 2-stage regression and MLR: *F* statistic

A simple null hypothesis of the form $H_0: \beta_1 = 0$ in MLR can be tested with the F-test (or equivalently with the t-test). The F-statistic can be expressed as

$$F = \frac{\rho^2_{YS|C}}{(1 - \rho^2_{YS|C})/(n-2)}$$

If the null hypothesis is true, then F follows an F-distribution with $(1, n-2)$ degrees of freedom, where n is the number of observations. Similarly, in a 2-stage residual outcome analysis $F^* = \frac{\rho^2_{(Y|C)S}}{(1 - \rho^2_{(Y|C)S})/(n-2)}$ follows an F-distribution with $(1, n-2)$ degrees of freedom under the null hypothesis $H_0 : \beta^*_1 = 0$

As in the case of the regression parameters, the F-statistics from MLR and 2-stage also have a simple form of relationship that can be given by $F^* = \frac{m - \rho^2_{SC}}{m} F$, where $m = 1 + \rho^2_{SC}\rho^2_{YS|C} - \rho^2_{YS|C}$. When $\rho^2_{SC} = 0$, the two approaches produce identical results. With increase in ρ^2_{SC}, however, the two F-statistics can lead to substantially different conclusions. For example, for a SNP effect of $\rho^2_{YS|C} = 1\%$ and $\rho^2_{SC} = 0, 0.2, 0.5, 0.8$, the correction factors would be $\frac{m - \rho^2_{SC}}{m} = 1.0$, 0.789, 0.497, and 0.198, respectively. For $\rho^2_{SC} = 0.2$, for example, the F-statistic in a 2-stage analysis will be attenuated approximately by 21%.

3. Example: Genome-wide association study of femoral neck length (FNL)

To illustrate the two approaches (MLR and 2-stage), we considered a genome-wide association study of femoral neck length (FNL, cm) in the Framingham Osteoporosis Study. FNL is one of several hip geometry phenotypes that have been implicated to play an important role in fracture risk [Kaptoge et al., 2008]. The design and methods of the Framingham Osteoporosis Study, an ancillary study of the Framingham Heart Study (FHS), have been described elsewhere in detail [Karasik et al., 2010, Hsu et al., 2010]. Briefly, members of the Framingham Osteoporosis Study are participants of the Original and Offspring cohorts of FHS who had scans of the femur measured by DXA with a Lunar DPX-L (Lunar Corp., Madison, WI, USA). Scans were performed between 1996 and 2001 for the Offspring cohort and 1992 and 1993 for Original cohort and a hip structure analysis program (HSA) [Khoo et al., 2005] was used to derive FNL measures. Information on age, sex, height, and body mass index (BMI) was obtained for each participant at the time of the DXA measurement. Genotyping was conducted by the Framingham SHARe (SNP Health Association Resource) project, for which 549,827 SNPs (Affymetrix 500K mapping array plus Affymetrix 50K gene center array) were genotyped. For genome-wide association analysis, a total of 2,543,887 autosomal SNPs were imputed by MACH2 (http://www.sph.umich.edu/csg/abecasis/MACH/) using HapMap phased haplotypes (release 22, build 26, CEU population) as a reference panel. A principal component analysis was performed to derive principal components of ancestry [Price et al., 2006]. Since our theoretical analyses assumed independence, we limited our analyses of FHS data here to 2173 (1235 women) *unrelated* subjects selected from the Original and Offspring cohorts who provided blood samples for DNA and had scans of the femur measured by DXA. The study was approved by the Institutional Review Boards for Human Subjects Research at Boston University and the Hebrew Rehabilitation Center.

Genome-wide association analysis of femoral neck length (FNL, cm) was conducted using MLR and 2-stage regression and adjusting for age, sex, height, BMI, and the first four principal components (PC1–PC4) of ancestry to account for confounding due to potential

population substructure. We analyzed 2,144,878 SNPs after excluding SNPs with low minor allele frequency (< 5%) and low imputation quality (<0.3) as measured by the ratio of the empirically observed dosage variance to the expected (binomial) dosage variance.

Consistent with our theoretical findings, p-values (and effect sizes) from the 2-stage analysis were weaker than those obtained from MLR. While the p-values of the 2-stage were systematically weaker than those of MLR, no substantial differences were observed for the association of any single SNP and FNL as associations between the covariates and SNPs across the genome were only minimal. Among SNPs with p<0.05, the strongest SNP-covariates association was observed for rs17615220 (squared multiple correlation of r^2) resulting in only minor differences between the MLR and the 2-stage results: the regression parameter estimate, t-statistic, and p-value for rs17615220 and FNL association were 0.081, 2.43, and 0.015 for MLR and 0.068, 2.24, and 0.025 for 2-stage.

To further illustrate differences between MLR and 2-stage in the presence of stronger correlations between independent variables, we also performed association analyses of FNL and arbitrarily selected SNPs with varying degrees of linkage disequilibrium (LD, r^2). This approach could occur in a conditional analysis, where residuals adjusting for known variants are generated for a two-stage approach. For illustrative purposes, we selected rs764300, the most significant SNP on chromosome 14, as the SNP of interest and performed association analysis with FNL conditional on covariates (age, sex, height, BMI, and PC1–PC4) only and then additionally adjusting for each of the following SNPs: rs4902067, rs12883544, rs912343, rs1016247, having LD with rs764300 ranging between r^2 =0.0 and 0.53 (Table 1). As expected, when r^2 is close to 0 (rs4902067), associations between the index SNP and FNL in MLR and 2-stage approach were nearly identical (*P-value*=3.0E-05, parameter estimate=0.09). On the other hand, when r^2 between the index SNP and the other SNP in the model is high (r^2 =0.53, rs1016247), the conditional association from the 2-stage was substantially different from MLR (2-stage *P-value* =0.0538, parameter estimate=0.042, and t-statistic=1.93; MLR *P-value* =0.0047, parameter estimate=0.089, and t-statistic=2.83). As shown in our theoretical presentation, the effect size (parameter estimate) from the 2-stage analysis was attenuated by 53%.

4. CONCLUSION

We examined the performance of a 2-stage residual-outcome analysis as compared with multiple linear regression (MLR) analysis and showed that the two approaches can produce vastly different results. Our results showed that compared to MLR a 2-stage residual-outcome analysis can markedly underestimate and fail to detect a SNP effect. Unlike MLR analysis which removes the contribution of a covariate from both the outcome and exposure variables, the 2-stage residual-outcome analysis removes a covariate effect *only* from the outcome variable. An important issue of whether a covariate effect should be adjusted only from the outcome (adjusted-outcome), only from the exposure (adjusted-exposure) or from both the outcome and exposure variables (adjusted-outcome and adjusted-exposure) must be determined according to the goal of the analysis. In doing so, however, it is critical to make distinctions between a 2-stage approach and MLR and select the appropriate method. This is especially important when a study involves correlated independent variables in which the two approaches are expected to be incongruent.

There are valid and useful applications to a 2-stage residual-based analysis. In nutritional epidemiologic studies, for example, residuals or energy-adjusted nutrient values are created prior to the main analysis to remove variations due to energy intake from nutrients [Willett WC; 1998]. This approach is used to obtain a measure of a nutrient intake independent of total energy to minimize issues related to collinearity, to obtain association estimates of total

energy that are not adjusted for specific nutrients and to reduce correlations among nutrients by removing co-variation due to shared measurement errors in self-reported nutrients and energy intake [Willett WC, 2001; Michels et al., 2004].

Residual-outcome (adjusted-outcome) analysis has been also very popular in genetic linkage and family–based association studies [Rabinowitz and Laird, 2000; Family-Based Association Tests and FBAT-toolkit, FBAT, http://www.biostat.harvard.edu/~fbat /fbat.htm; Lunetta et al., 2000; Slager et al. 2003; Zeegers et al. 2004]. Based on empirical data, many studies indicate findings from 2-stage residual-outcome analysis to be comparable to those obtained from a single-stage method that jointly analyze the study variables. For sib-pair-linkage studies, a simulation study by Zeegers et al. demonstrated that a residual-outcome analysis performs as well as a joint analysis of an outcome, a quantitative trait locus and covariates. In general, however, little is known about how well a 2-stage residual-outcome analysis performs under various scenarios and different study designs.

In this report our focus was on a 2-stage residual-outcome approach for analysis of potential confounding effects, SNPs, and quantitative outcome variables in a population-based association study. We showed, using theoretical analysis and empirical data, that the treatment of confounding effects by a 2-stage residual-outcome analysis is inconsistent with that of multiple linear regression analysis.

Acknowledgments

REFERENCES

Christenfeld N, Sloan R, Carroll D, Greenland S. Risk Factors, Confounding, and the Illusion of Statistical Control. Psychosomatic Medicine. 2004; 66:868–875. [PubMed: 15564351]

Family-Based Association Tests and FBAT-toolkit (user's manual. 2009 March. http://www.biostat.harvard.edu/~fbat/fbat.htm

Hennekens, CH.; Buring, JE.; Mayrent, SH. Epidemiology in Medicine. Boston: Little, Brown; 1987.

Hsu YH, Zillikens MC, Wilson SG, Farber CR, Demissie S, Soranzo N, Bianchi EN, Grundberg E, Liang L, Richards JB, Estrada K, Zhou Y, van Nas A, Moffatt MF, Zhai G, Hofman A, van Meurs JB, Pols HA, Price RI, Nilsson O, Pastinen T, Cupples LA, Lusis AJ, Schadt EE, Ferrari S, Uitterlinden AG, Rivadeneira F, Spector TD, Karasik D, Kiel DP. An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits. PLoS Genet. 2010; 6:e1000977. [PubMed: 20548944]

Kaptoge S, Beck TJ, Reeve J, Stone KL, Hillier TA, Cauley JA, Cummings SR. Prediction of incident hip fracture risk by femur geometry variables measured by hip structural analysis in the study of osteoporotic fractures. J Bone Miner Res. 2008; 23(12):1892–1904. [PubMed: 18684092]

Karasik D, Dupuis J, Cho K, Cupples LA, Zhou Y, Kiel DP, Demissie S. Refined QTLs of osteoporosis-related traits by linkage analysis with genome-wide SNPs: Framingham SHARe. Bone. 2010; 46(4):1114–1121. [PubMed: 20064633]

Khoo BC, Beck TJ, Qiao QH, Parakh P, Semanick L, Prince RL, Singer KP, Price RI. In vivo short-term precision of hip structure analysis variables in comparison with bone mineral density using paired dual-energy X-ray absorptiometry scans from multi-center clinical trials. Bone. 2005; 37:112–121. [PubMed: 15869917]

Kleinbaum; Kupper; Muller. Applied Regression Analysis and Other Multivariable Methods. 2nd Edition. Duxbury Press; 1988.

Lunetta KL, Farone SV, Biederman J, Laird NM. Family based tests of association and linkage using unaffected sibs, covariates and interactions. Am J Hum Gen. 2000; 66:605–614.

Michels KB, Bingham SA, Luben R, Welch AA, Day NE. The Effect of Correlated Measurement Error in Multivariate Models of Diet. Am J Epidemiology. 2004; 160:59–67.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8): 904–909. [PubMed: 16862161]

Rabinowitz D, Laird NM. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered. 2000; 50:211–223. [PubMed: 10782012]

Rothman, KJ.; Greenland, S. Modern Epidemiology. 2nd Ed.. 1998.

Slager SL, Iturria SJ. Residuals in defining phenotypes – 2-stage approach to define phenotypes for linkage study: Genome-wide linkage analysis of systolic blood pressure: a comparison of two approaches to phenotype definition. BMC Genetics. 2003; 4 Suppl 1

Willett, WC. Nutritional Epidemiology. 2nd Edition. New York: Oxford University Press; 1998.

Willett WC. Dietary diaries versus food frequency questionnaires—a case of undigestible data. Int J Epidemiol. 2001; 30:317–319. [PubMed: 11369736]

Zeegers M, Rijsdijk F, Sham P. Adjusting for Covariates in Variance Components QTL Linkage Analysis. Behavior Genetics. 2004; 34(2):127–133. [PubMed: 14755177]

Appendix

Relationship between 2-stage regression and MLR parameter estimates

This section uses classical linear regression and correlation theories (e.g., Kleinbaum et al., 1988) to derive a simple relationship between regression parameters from MLR and 2-stage regression models, β_1 and β_1^*, and show that $\beta_1^* = \beta_1(1 - \rho_{SC}^2)$ where ρ_{SC}^2 the squared correlation between S and C.

From Eq. (3), we have $\beta_1^* = \rho_{ZS} \dfrac{\sigma_Z}{\sigma_S}$

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

Where $\rho_{YS|C}$ is the *partial* correlation between Y and S (both adjusted for C) and ρ_{SC}^2 the squared correlation between S and C

In addition, the variance of S can be given by $\sigma_S^2 = \dfrac{\sigma_{S|C}^2}{(1 - \rho_{SC}^2)}$

Thus,

$$\beta_1^* = \rho_{YS|C}\sqrt{1 - \rho_{SC}^2}\, \frac{\sigma_{Y|C}}{\sqrt{\frac{\sigma_{S|C}^2}{(1-\rho_{SC}^2)}}} = \rho_{YS|C}\sqrt{1 - \rho_{SC}^2}\,\sigma_{Y|C}\frac{\sqrt{1-\rho_{SC}^2}}{\sigma_{S|C}} = \rho_{YS|C}\frac{\sigma_{Y|C}}{\sigma_{S|C}}(1 - \rho_{SC}^2)$$

$$= \beta_1(1 - \rho_{SC}^2)$$

Table 1

Association between rs764300 and femoral neck length (FNL) using MLR and 2-stage analyses

Covariates	*r^2*	Multiple Regression			2-Stage Regression		
		Estimate	SE	P-value	Estimate	SE	P-value
Age, Sex, Height, BMI, PC1-PC4	-	0.090	0.022	0.00004	0.0890	0.022	0.00004
Age,Sex,Height, BMI, PC1-PC4, rs4902067	0.00	0.090	0.022	0.00003	0.0890	0.021	0.00003
Age,Sex,Height, BMI, PC1-PC4, rs12883544	0.16	0.080	0.024	0.00074	0.0670	0.022	0.00199
Age,Sex,Height, BMI, PC1-PC4, rs912343	0.31	0.077	0.026	0.00329	0.0530	0.022	0.01492
Age,Sex,Height, BMI, PC1-PC4, rs1016247	0.53	0.089	0.032	0.00479	0.0420	0.022	0.05385

*
linkage disequilibrium between rs764300 and the SNP included as a covariate in the model