A Comparison of Approaches to Control for Confounding Factors by Regression Models

Guan Xing^a Chang-Yun Lin^b Chao Xing^{b, c}

^aBristol-Myers Squibb Company, Pennington, N.J., ^bMcDermott Center of Human Growth and Development and

^cDepartment of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Tex., USA

Key Words

Linear regression • Confounding factor • Adjustment

Abstract

A common technique to control for confounding factors in practice is by regression adjustment. There are various versions of regression modeling in the literature, and in this paper we considered four approaches often seen in genetic association studies. We carried out both analytical and simulation studies comparing the bias of effect size estimates and examining the test sizes under the null hypothesis of no association between an outcome and an exposure. Further, we compared the methods in a nonsynonymous genome-wide scan for plasma lipoprotein(a) levels using a dataset from the Dallas Heart Study. We found that a widely employed approach that models the covariate-adjusted outcome and the exposure leads to an infranominal test size and underestimation of the exposure effect size. In conclusion, we recommend either using multiple regression models or modeling the covariate-adjusted outcome and the covariate-adjusted exposure to control for confounding factors.

Copyright © 2011 S. Karger AG, Basel

Introduction

When investigating association between an exposure and an outcome, it is important to examine whether the association is due to the effects of confounding factors. A factor is considered to be a potential confounder when it meets 3 basic conditions: (i) it must be associated with the outcome; (ii) it must be associated with the exposure; and (iii) it must not be an intermediate variable in the causal relationship between the exposure and the outcome [1]. Depending on the direction and magnitude of the relationships a confounding factor has with the exposure and the outcome, it can lead to either overestimation or underestimation of the true association. Thus, when it is not feasible or possible to prevent confounding at the study design stage, e.g. in non-randomized observational studies, proper strategies must be taken to control for the confounding factors at the data analysis stage. There are 3 types of techniques commonly employed in practice. One is stratification on levels of the confounding factors such that association is assessed within distinct groups of relatively homogeneous subjects. Alternatively, we can incorporate the exposure, the outcome, and any potential confounding factors into regression models such that asso-

ciation is assessed at fixed values of the confounding factors. Another method is propensity score adjustment [2, 3], which balances the exposure groups with regard to the confounding factors.

In genetic association studies, a salient source of confounding is population stratification – systematic difference in allele frequencies between subpopulations due to different ancestry. In the study design respect, a family-based study using the transmission-disequilibrium test [4] can avoid the confounding effects of population stratification. In the analysis respect, there are mainly 2 types of strategies to control for population stratification. One is called genomic control method, which uses multiple independent marker loci to adjust the distribution of a standard test statistic [5]; the other is referred to as structured association method, which infers proxies of ancestry and controls for them in association analysis [6–8]. In the genome-wide association (GWA) study era, the structured association method is popular because ancestral information can be accurately inferred using genome-wide markers. All 3 techniques – stratification, regression adjustment, and propensity score adjustment – have been applied in GWA studies [7, 9–11], and regression adjustment is most dominant because of, presumably, its simplicity in implementation.

In the literature, there are mainly 4 versions of regression modeling to adjust for population stratification. We describe in the Method section each approach accompanied by a reference paper using it. However, it is unclear whether all 4 approaches are valid in controlling for the confounding factors, and there has been no study comparing their power. In this study, we performed theoretical analysis comparing the 4 approaches by linear regression models, and carried out simulation studies to confirm and elaborate the analytical results. We further compared the 4 approaches in a nonsynonymous genome-wide scan for plasma lipoprotein(a) [Lp(a)] levels from the Dallas Heart Study (DHS). In particular, we compared the 4 approaches when multiple risk factors were present in addition to the confounding factors, which is the norm in real data analysis.

Method

Denote by Y the outcome, by G the exposure, and by X a vector of potential confounding factors and risk factors. In a genetic scenario, Y stands for the trait of interest, G is the genotype, and X can be confounding factors such as ancestry. Suppose the primary object is to investigate whether Y and G are correlated while taking into account the confounding effect of X . There are various

ways to achieve this goal, and below we describe 4 regression models commonly seen in the literature. For simplicity, assume Y and G are continuous variables and $\text{Var}(Y) = \sigma^2$. A natural way to account for the confounding factors in investigating the association between Y and G is by a multiple linear regression model: $Y = X\beta + G\gamma_1 + \varepsilon_{Y|X,G}$ (a), where β and γ_1 are regression coefficients, and $\varepsilon_{Y|X,G}$ is the random error term with $E(\varepsilon_{Y|X,G}) = 0$ and $\text{Var}(\varepsilon_{Y|X,G}) = \sigma^2$. Let the first column of X be $J = (1, 1, \dots, 1)'$ to incorporate an intercept term in the model. We denote this modeling as approach I [12]. There are various two-step approaches to adjust for the confounding factors. One is that, first, Y and G are separately regressed on X , i.e. $Y = X\beta + \varepsilon_{Y|X}$ (b), and $G = X\beta + \varepsilon_{G|X}$ (c), and then a regression model is fit to the residuals – $\varepsilon_{Y|X} = \gamma_{II0} + \gamma_{II1}\varepsilon_{G|X} + \varepsilon_{II}$ (d). We denote this modeling as approach II [7]. Another widely employed two-step procedure, which we denote as approach III [13], is to regress the residuals $\varepsilon_{Y|X}$ from model b on G by $\varepsilon_{Y|X} = \gamma_{III0} + \gamma_{III1}G + \varepsilon_{III}$ (e). A relatively less taken two-step procedure, which we denote as approach IV [14], is to regress Y on the residuals $\varepsilon_{G|X}$ from model c by $Y = \gamma_{IV0} + \gamma_{IV1}\varepsilon_{G|X} + \varepsilon_{IV}$ (f). In all 4 approaches, one tests the hypotheses $H_0: \gamma_i = 0$ versus $H_1: \gamma_i \neq 0$, where $i \in \{I, II, III, IV\}$, to determine whether Y and G are associated while controlling for the confounding effect of X . Since all 4 approaches are employed in the literature, it is of interest to compare them by investigating the regression coefficient estimates $\hat{\gamma}_i$ in terms of bias and examining whether conventional tests have a proper size.

First, we carried out theoretical analysis without regard for the actual relationships among Y , G , and X . However, in practice it is of particular interest to compare the performance of the 4 approaches when there is no relationship between Y and G , but there exist confounding factors X complicating their relationship. We carried out 2 sets of simulation under this condition. In the first set, we performed proof-of-principle simulation across the parameter space of the correlation coefficients between X and G , and between Y and X by the following strategy. First, (X, G) were simulated assuming a bivariate normal distribution with the correlation coefficient $\rho_{X,G}$ varying from -0.9 to 0.9 in increments of 0.1 . Second, Y was simulated according to a linear model $Y = \beta X + \varepsilon$, in which β and ε were chosen such that $\rho_{Y,X}$ varied from -0.9 to 0.9 in increments of 0.1 . Note that when $\rho_{X,G} = 0$ and/or $\rho_{Y,X} = 0$, X did not act as a confounding factor, whereas in all other cases X was a confounding factor. The sample sizes were set over the range $\{30, 100, 500\}$. 10,000 replicate samples were generated under each simulation condition. We analyzed the simulated data by approaches I–IV as well as a regression model without adjusting for X , which we denoted as approach V.

In the second set of simulation, based on the 1,000 Genomes Project data [15], we generated genotypic data G spanning a region of $\sim 1,000,000$ base pairs centered around the apo(a) gene on chromosome 6 to mimic a real genetic study. In particular, we simulated 5,000 individuals based on the haplotypes of 87 individuals from the Centre d'Etude du Polymorphisme Humain collected in Utah, USA, with ancestry from northern and western Europe (CEU), and simulated another 5,000 individuals based on the haplotypes of 88 individuals from Yoruba in Ibadan, Nigeria (YRI) using the software HAPGEN2 [16]. We simulated normally distributed traits Y in the CEU- and YRI-derived samples by assuming an equal variance in both populations and varying the difference of trait means by d , $d \in \{0.0, 0.1, 0.2, \dots, 0.9, 1.0\}$, times the standard deviation of the trait. In each of the 11 sets of pheno-

type configuration 1,000 replicates were generated. Suppose these 10,000 individuals would be grouped together for analysis with a binary variable X indicating each person's ethnicity. Again we analyzed the simulated data by approaches I–V.

We further compared the 4 approaches in a GWA study of plasma Lp(a) levels from the DHS. Lp(a) levels vary widely between populations, with African Americans (AAs) having much higher levels than European Americans [17]. A genome-wide scan without adjusting for ancestry will presumably generate many spurious associations in an AA sample. Thus this provides us with an opportunity to evaluate the 4 approaches to control for confounding factors, in particular ancestry, in GWA studies. The majority of the variation in Lp(a) levels is accounted for by the apo(a) gene [18]. A kringle IV-encoding region in the gene determines the size of apo(a) isoforms, which is inversely correlated with Lp(a) levels [19]. Thus this further provides us with an opportunity to evaluate the 4 approaches to adjust for risk factors in addition to confounders in GWA studies.

The DHS is a multiethnic, population-based cohort in Dallas County [20]. In this study we focused on the AAs, in which 1,726 individuals genotyped with 8,968 nonsynonymous single-nucleotide polymorphisms (SNPs) and 2,270 ancestry-informative SNPs [21] across the autosomes. We filtered out those with minor allele frequency less than 0.01 and those out of Hardy-Weinberg equilibrium ($p < 1.0 \times 10^{-5}$), resulting in 10,686 SNPs for further analysis. Two sets of genome-wide scans were carried out using the 4 approaches. First, we only included ancestry, which was inferred for each individual using ancestry-informative markers as described elsewhere [22], as a covariate. Second, we included apo(a) isoform size, age and sex besides ancestry as covariates. In all analyses, Lp(a) levels were log-transformed and genotypic values were coded in an additive manner. As a comparison, we also performed a genome-wide scan without adjusting for any covariate, i.e. by approach V. The genomic inflation factor (GIF) [5], which is expected to be one without systemic deviation from the null hypothesis of no association genome-wide, was calculated for each analysis.

Results

The analytical details are described in the Appendix section, and here we summarize the theoretical results as well as simulation results. Approaches I, II, and IV yield the same estimate

$$\hat{\gamma}_I = \hat{\gamma}_{II} = \hat{\gamma}_{IV} = \frac{G'R_x Y}{G'R_x G}$$

and approach III yields a different estimate

$$\hat{\gamma}_{III} = \frac{G'R_x Y}{G'G - (\sum G)^2/n},$$

where $R_x = I_n - X(X'X)^{-1}X'$. There exists an inequality $|\hat{\gamma}_{III}| < |\hat{\gamma}_I = \hat{\gamma}_{II} = \hat{\gamma}_{IV}|$, because $G'G - (\sum G)^2/n$ and $G'R_x G$ correspond to the total sum of squares (SSTO) and the error sum of squares (SSE), respectively, in the regression

model c and we know $SSTO > SSE$. Assuming model a to be the underlying model, approaches I, II, and IV yield an unbiased estimate of the effect size of G , whereas approach III underestimates the effect size. The relative bias equals the coefficient of determination of model c. Therefore, the higher the correlation between X and G is, the more the effect size of G is underestimated by approach III, as confirmed by the simulation study (fig. 1).

There is a quantitative relationship between the variance estimates of the regression coefficients by approach I and approach II

$$\frac{Var(\hat{\gamma}_I)}{Var(\hat{\gamma}_{II})} = \frac{n-2}{n-(p+1)},$$

where p is the total number of independent variables in model a excluding J . Therefore, when $n \gg p$, these two variance estimates are very close. In other words, the asymptotic relative efficiency (ARE) of $\hat{\gamma}_I$ to $\hat{\gamma}_{II}$ equals one. There exists an inequality $Var(\hat{\gamma}_{IV}) > Var(\hat{\gamma}_{II})$, because

$$Var(\hat{\gamma}_{IV}) - Var(\hat{\gamma}_{II}) = \frac{SSTO - SSE}{(n-2)G'R_x G},$$

where SSTO and SSE apply to regression model b. Therefore, $ARE(\hat{\gamma}_{II}, \hat{\gamma}_{IV}) > 1$, and the higher the correlation between Y and X is, the less efficient $\hat{\gamma}_{IV}$ is compared to $\hat{\gamma}_{II}$ (and $\hat{\gamma}_I$, too), as confirmed by the simulation study (fig. 2).

Suppose we use Wald's test to test the hypotheses $H_0: \gamma_i = 0$ versus $H_1: \gamma_i \neq 0$. Wald's test statistic is defined as $W_i = \hat{\gamma}_i^2 / Var(\hat{\gamma}_i)$. Under H_0 , W_i is expected to follow a χ^2 distribution with 1 degree of freedom. W_I and W_{II} have similar asymptotic properties as

$$\lim_{n \rightarrow \infty} \frac{W_I}{W_{II}} = 1.$$

On the contrary, $W_{II} > W_{IV}$ because $\hat{\gamma}_{II} = \hat{\gamma}_{IV}$ and $Var(\hat{\gamma}_{IV}) > Var(\hat{\gamma}_{II})$. Therefore, a test based on W_{II} is more powerful than that based on W_{IV} , and the higher the correlation between Y and X is, the larger the power difference is. There exists an equality $W_{II} > W_{III}$ because

$$\frac{W_{II}}{W_{III}} = \frac{SSTO(Y'R_x Y) - (G'R_x Y)^2}{SSE(Y'R_x Y) - (G'R_x Y)^2},$$

where SSTO and SSE apply to regression model c. Therefore, a test based on W_{II} is more powerful than that based on W_{III} , and the higher the correlation between G and X is, the larger the power difference is. The relationships between W_{II} and W_{IV} , and between W_{II} and W_{III} are confirmed by the simulation study (fig. 3).

In the simulation study, we specifically investigated the validity of different approaches under the null hy-

Fig. 1. Relative bias of regression coefficient estimate by approach III. The relative bias is defined as $(\hat{\gamma}_i - \hat{\gamma}_{III})/\hat{\gamma}_i$, where $i \in \{I, II, IV\}$. The mean relative bias in the z-axis is based on 10,000 replicates with a sample size of 500 in each dataset.

Fig. 2. Difference of the standard deviation estimate of the regression coefficient by approach IV compared with that by approach II. The relative difference is defined as $(\sqrt{\text{Var}(\hat{\gamma}_{IV})} - \sqrt{\text{Var}(\hat{\gamma}_{II})})/\sqrt{\text{Var}(\hat{\gamma}_{II})}$. The mean relative difference in the z-axis is based on 10,000 replicates with a sample size of 500 in each dataset.

Table 1. Empirical type I error rates^a of different approaches at a nominal 0.05 significance level under various correlations among Y, X, and G, while Y and G are uncorrelated

$\rho_{Y,X}$	$\rho_{X,G} = -0.9$			$\rho_{X,G} = -0.6$			$\rho_{X,G} = -0.3$			$\rho_{X,G} = 0.0$		
	I	II	V	I	II	V	I	II	V	I	II	V
Sample size = 30												
-0.9	0.049	0.053	1.000	0.048	0.053	0.895	0.045	0.050	0.298	0.052	0.057	0.054
-0.6	0.050	0.055	0.891	0.052	0.056	0.513	0.050	0.054	0.165	0.051	0.055	0.052
-0.3	0.054	0.058	0.308	0.056	0.061	0.160	0.050	0.054	0.078	0.051	0.057	0.054
0.0	0.050	0.055	0.051	0.054	0.058	0.049	0.047	0.052	0.047	0.050	0.054	0.049
Sample size = 100												
-0.9	0.050	0.051	1.000	0.052	0.053	1.000	0.049	0.050	0.781	0.055	0.057	0.052
-0.6	0.051	0.053	1.000	0.050	0.051	0.959	0.049	0.050	0.437	0.050	0.052	0.052
-0.3	0.051	0.052	0.784	0.050	0.051	0.437	0.046	0.048	0.140	0.049	0.051	0.049
0.0	0.049	0.050	0.051	0.045	0.046	0.046	0.044	0.045	0.045	0.050	0.051	0.049
Sample size = 500												
-0.9	0.051	0.052	1.000	0.048	0.048	1.000	0.051	0.051	1.000	0.050	0.050	0.050
-0.6	0.048	0.048	1.000	0.051	0.052	1.000	0.053	0.053	0.984	0.052	0.052	0.050
-0.3	0.048	0.048	1.000	0.051	0.051	0.982	0.050	0.050	0.524	0.050	0.050	0.050
0.0	0.053	0.054	0.049	0.051	0.051	0.052	0.052	0.053	0.051	0.051	0.052	0.053

^a Based on 10,000 replicates.

Fig. 3. Comparison of approaches III and IV under various correlations among Y, X, and G, while Y and G are uncorrelated. **a** and **b** correspond to the difference of Wald's test statistics by approaches III and IV, respectively, compared with that by approach II. The relative difference is defined as $(W_{II} - W_i)/W_{II}$, where i denotes III and IV. **c** and **d** correspond to the empirical type I error rates of approaches III and IV, respectively, at a nominal 0.05 significance level. The results are based on 10,000 replicates with a sample size of 500 in each dataset.

Table 2. Comparison of empirical type I error rates and relative bias of effect size estimates by different approaches in genetic association studies with ancestry as a confounding factor

$\bar{\rho}_{Y,X}^a$	$ \rho_{X,G} ^b$	Mean type I error rate ^c					Mean relative bias ^d
		I	II	III	IV	V	
0.00	[0.0, 0.1)	0.049	0.049	0.049	0.049	0.049	0.002
	[0.1, 0.2)	0.050	0.050	0.048	0.050	0.050	0.026
	[0.2, 0.3)	0.051	0.051	0.046	0.051	0.050	0.053
	[0.3, 0.4)	0.051	0.051	0.042	0.051	0.049	0.097
	[0.4, 0.5)	0.049	0.049	0.029	0.049	0.050	0.210
0.10	[0.0, 0.1)	0.049	0.049	0.049	0.048	0.076	0.001
	[0.1, 0.2)	0.050	0.050	0.048	0.049	0.294	0.023
	[0.2, 0.3)	0.051	0.051	0.044	0.051	0.629	0.050
	[0.3, 0.4)	0.050	0.050	0.040	0.049	0.876	0.096
	[0.4, 0.5)	0.050	0.050	0.025	0.048	1.000	0.214
0.20	[0.0, 0.1)	0.049	0.049	0.049	0.046	0.158	0.002
	[0.1, 0.2)	0.050	0.050	0.048	0.045	0.759	0.025
	[0.2, 0.3)	0.050	0.050	0.046	0.047	0.992	0.050
	[0.3, 0.4)	0.051	0.051	0.045	0.047	1.000	0.096
	[0.4, 0.5)	0.048	0.048	0.028	0.045	1.000	0.213
0.29	[0.0, 0.1)	0.049	0.049	0.049	0.040	0.270	0.001
	[0.1, 0.2)	0.050	0.050	0.048	0.041	0.957	0.019
	[0.2, 0.3)	0.051	0.051	0.046	0.042	1.000	0.053
	[0.3, 0.4)	0.050	0.050	0.040	0.043	1.000	0.095
	[0.4, 0.5)	0.051	0.051	0.031	0.041	1.000	0.210
0.41	[0.0, 0.1)	0.049	0.049	0.049	0.031	0.417	0.001
	[0.1, 0.2)	0.050	0.050	0.048	0.031	0.999	0.024
	[0.2, 0.3)	0.050	0.050	0.045	0.034	1.000	0.053
	[0.3, 0.4)	0.050	0.050	0.041	0.035	1.000	0.097
	[0.4, 0.5)	0.049	0.049	0.030	0.030	1.000	0.208

^a Mean correlation coefficient between ancestry and a trait based on 1,000 replicates. Assuming an equal variance for the trait in the two populations, $\bar{\rho}_{Y,X} = 0.00, 0.10, 0.20, 0.29$ and 0.41 correspond to the difference of trait means equal to $d = 0.0, 0.2, 0.4, 0.6$ and 0.9 , times the standard deviation of the trait, respectively.

^b Absolute value of correlation coefficients between ancestry and genotype, according to which the 4,744 SNPs were grouped into five groups.

^c Mean of the empirical type I error rates, which were calculated at a nominal 0.05 significance level based on 1,000 replicates for each SNP, in a SNP group defined by $\rho_{X,G}$.

^d The relative bias was defined as $(\hat{\gamma}_i - \hat{\gamma}_{III})/\hat{\gamma}_i$, where $i \in \{I, II, IV\}$. We first calculated the mean relative bias for each SNP based on 1,000 replicates, and then took the mean of all SNPs in a group.

pothesis of no relationship between Y and G . When X is a confounding factor, i.e. neither $\rho_{X,G}$ nor $\rho_{Y,X}$ equals zero, a test without adjusting for it – approach V – results in an inflated type I error rate, and the degree of inflation is positively correlated with the degree of correlation. When the sample size is small ($n = 30$), approach I shows a proper size whereas approach II has a supranominal size with the type I error rate inflated by $\sim 10\%$ compared to

that of approach I (table 1). By a one-sided two-sample proportion test on the empirical data of 10,000 replicates in each scenario, we observe that when the difference between the two type I error rates is less than ~ 0.0055 , it is not statistically significant at a nominal level of 0.05; when the difference is greater than ~ 0.0055 , it is significant at the level of 0.05, but insignificant at the level of 0.01. When the sample size is moderate to large, both ap-

Fig. 4. Relative bias of variants effect size estimates by approach III on Lp(a) levels in the DHS. **a** and **b** correspond to models adjusting for ancestry only and models adjusting for four covariates – ancestry, apo(a) isoform size, age and sex – respectively. The relative bias is defined as $(\hat{\gamma}_i - \hat{\gamma}_{III})/\hat{\gamma}_i$, where $i \in \{I, II, IV\}$ and the mean relative bias is based on each quarter of SNPs categorized by the correlation coefficients between ancestry and genotype. The ranges of the correlation coefficients for each quarter are (0, 0.04), (0.04, 0.08), (0.08, 0.15), and (0.15, 0.34).

proaches I and II maintain a proper size regardless of $\rho_{X,G}$ and $\rho_{Y,X}$. On the contrary, approach III shows an infranominal size when $|\rho_{X,G}| > 0$ regardless of $\rho_{Y,X}$, and approach IV shows an infranominal size when $|\rho_{Y,X}| > 0$ regardless of $\rho_{X,G}$. The higher the correlation between G and X is, the smaller the test size of approach III is; the higher the correlation between Y and X is, the smaller the test size of approach IV is. Approaches III and IV maintain a proper size only when $\rho_{X,G} = 0$ and $\rho_{Y,X} = 0$, respectively (fig. 3).

In the simulation mimicking a genetic study testing for association between a phenotype and SNPs in mixed populations, we investigated validity of different approaches under H_0 and the bias of effect size estimates by approach III (table 2). There were a total of 14,278 SNPs called in the region from the whole 1,000 Genomes Project data, among which 4,744 were present in both CEU and YRI. Analyses were focused on these SNPs. The correlation coefficients between ancestry and SNPs ranged from 0.00 to 0.49, and the correlation coefficients between ancestry and the simulated traits ranged from 0.00

to 0.45. Consistent with the analytical and proof-of-principle simulation results, approaches I and II maintained proper type I error rates, whereas approaches III and IV resulted in deflated type I error rates when $|\rho_{X,G}| > 0$ and $|\rho_{Y,X}| > 0$, respectively, though the deflation was negligible when $|\rho| < 0.1$. Approach III led to underestimated effect sizes with a relative bias $> 10\%$ when $|\rho_{X,G}| \geq 0.3$.

In the DHS AAs, the mean proportion of European ancestry was 0.22 with a standard deviation of 0.08. Lp(a) levels were significantly associated with ancestry as well as apo(a) isoform size, age and sex ($p = 3.8 \times 10^{-5}$, 1.3×10^{-257} , 3.5×10^{-3} , and 2.5×10^{-6} , respectively). Isoform size explained 40% of variance in Lp(a) levels, whereas ancestry only explained 0.9%. These four variables together accounted for 45% of the total variance. The most significant SNP associated with Lp(a) levels was rs1801693 located in apo(a), with p values equal to 1.1×10^{-12} and 7.8×10^{-4} for analyses by approach I adjusting for only ancestry and adjusting for all four variables, respectively. It showed an ethnic disparity with the minor allele frequencies being 0.13 and 0.31 in DHS AAs and European Americans, respectively; however, its correlation coefficient with ancestry was only 0.08, and 0.18 with isoform size. For each SNP we calculated the relative bias of the regression coefficient estimate by approach III compared with that by other approaches, and reported the mean relative bias for each quarter of SNPs based on the correlation coefficients between ancestry and genome-wide SNPs, which ranged from 0.00 to 0.34. Consistent with the analytical and simulation results, the higher the correlation between a SNP and ancestry was, the more the effect size of a variant was underestimated by approach III (fig. 4).

The quantile-quantile plots for each analysis were illustrated in figure 5. Without adjusting for any covariate, the GIF is 1.165, clearly showing a systemic deviation from H_0 genome-widely. Approaches I–IV adjusting for only ancestry lead to GIFs of 1.018, 1.019, 1.007 and 1.009, respectively, and adjusting for all four variables lead to GIFs of 1.063, 1.065, 1.056, and 0.620, respectively, confirming ancestry as a confounding factor. Consistent with the theoretical analysis, no matter how many covariates were adjusted for, approaches I and II generated similar results with the latter being slightly more liberal, and approaches III and IV were conservative compared with approaches I and II. Both approaches I and II effectively controlled for the population stratification.

Fig. 5. Quantile-quantile plots of genome-wide scans for Lp(a) levels in the DHS. Columns I–V correspond to approaches I–V, respectively. The first row corresponds to analyses with ancestry as a single covariate, and the second row corresponds to analyses with ancestry, apo(a) isoform size, age and sex as covariates.

Discussion

Confounding factors distort the true relationship between an exposure and an outcome. Thus one needs to properly control for them to obtain an unbiased estimate of the exposure effect. In this paper we compared 4 linear regression models, and showed that approach III resulted in underestimated effect sizes, whereas the other 3 provided an unbiased estimate. Note that this conclusion holds in general, not only in adjusting for confounding factors, but also in adjusting for non-confounding risk factors, though the bias will be minimal when a risk factor is independent of the exposure. The fact that approaches I and II yield identical estimates is relatively well-known, and it actually forms the basis for algorithms in multiple regression [23]. In analyzing age-adjusted rates data, Rosenbaum and Rubin [24] discussed another model – regressing the residuals $\varepsilon_{Y|X}$ from model b on both G and X . By replacing Y with $R_X Y$, it is easy to show this model provides an unbiased estimate, too. In this study we did not discuss this model because it is usually constrained to studies where only covariate-adjusted outcomes are obtainable.

In adjusting for covariates, approach III has an infranominal test size when the covariates and the expo-

sure are correlated, and approach IV has an infranominal test size when covariates and the outcome are correlated. There is an intuitive explanation for this phenomenon. Suppose a covariate is correlated with the exposure. When the covariate is regressed on the outcome by approach III, part of the information on the exposure is also regressed out, which weakens the association between the covariate-adjusted residual and the exposure. In an extreme case of the covariate being identical to the exposure, the covariate-adjusted residual is orthogonal to the exposure. Similarly, arguments can be made for approach IV. Following this reasoning, our conclusion should hold not only for linear regression models, but also other methods with analogous logic, though it is difficult to derive a closed-form solution in some cases. Note also that, when the exposure and the outcome are clearly defined in terms of a possible causality relationship, approaches III and IV are distinctive; however, when the identities of exposure and outcome are ambiguous, these two approaches are exchangeable.

In this paper we have been considering Wald's test for hypothesis testing, and show that there exist relationships

$$\lim_{n \rightarrow \infty} \frac{W_I}{W_{II}} = 1,$$

$W_{II} > W_{III}$ and $W_{II} > W_{IV}$. The choice of Wald's test is out of demonstrative purpose because its test statistic can be dissected into 2 parts: a coefficient estimate and its variance estimate. It particularly helps elucidate the differences between approaches when the same coefficient estimate is reached by approaches I, II, and IV. Besides Wald's test, there are 2 other tests commonly employed – the likelihood ratio test and the score test. Denote their test statistics by LR and S, respectively. Although the 3 tests have equivalent asymptotic power for testing local alternatives [25], there exists a systemic inequality $W \geq LR \geq S$, and in linear regression models, it can be derived that $LR = n \log(1 + W/n)$ and $S = W/(1 + W/n)$ [26]. Based on these 2 equations, it can be shown that

$$\lim_{n \rightarrow \infty} \frac{LR_I}{LR_{II}} = \lim_{n \rightarrow \infty} \frac{S_I}{S_{II}} = 1,$$

and, given $W_{II} > W_j$, where $j \in \{III, IV\}$, there also exist inequalities $LR_{II} > LR_j$ and $S_{II} > S_j$. Therefore, the conclusions we draw on the test size of the 4 approaches based on Wald's test apply to the likelihood ratio test and the score test, too.

In the GWA scan for plasma Lp(a) levels in DHS, when only ancestry was taken into account, both approaches III and IV generated GIFs (1.007 and 1.009, respectively) close to, though greater than, 1 implying that both approaches also maintained a proper size. The GIFs were only marginally smaller than those of approaches I and II (1.018 and 1.019, respectively). This phenomenon could be explained by ancestry's correlations with SNPs and Lp(a) levels – the correlation coefficients between ancestry and SNPs were low, even for those ancestry-informative ones such as rs1801693 ($\rho = 0.08$); similarly, ancestry only explained 0.9% of variance in Lp(a) levels. However, these 2 approaches behaved differently when all four covariates were taken into account. Approach III still retained a proper size because the extra covariates should be uncorrelated with SNPs genome-widely. On the contrary, approach IV had an infranominal size because the four covariates together accounted for 45% of variance in Lp(a) levels. When regressing them on the genotype, much of the information on Lp(a) levels was regressed out too, which consequently led to dissociation between covariate-adjusted genotype and Lp(a) levels.

Different from approach IV, in approaches I, II, and III, the GIF was higher when all four covariates were adjusted for than that when only ancestry was taken into account. We speculate the underlying statistical reason is that omitting covariates associated with the response

results in estimation efficiency loss for the included variables in generalized linear models [27]. We had showed that in testing the null hypothesis of no genetic effect, omitting disease risk/preventive factors always leads to efficiency loss, and the magnitude of loss increases with the strength of the association between the omitted covariate and the disease [28]. In the current study, apo(a) isoform size, age and sex accounted for >40% of variance in Lp(a) levels. Therefore, omitting them led to power loss in testing for association between phenotype and genotype, which was reflected in the decrease of GIFs.

Special attentions should be paid to approach III, a very popular technique in practice, in particular, when methods that cannot directly handle covariates are chosen to test association between the outcome and the exposure. When only risk factors independent of the exposure of primary interest are to be adjusted for, this approach is valid; however, when potential confounding factors – which are correlated with the exposure by definition – are to be controlled for, this approach leads to power loss in detecting association and underestimation of the exposure effect size in characterizing the association. This approach is widely applied in genetic association studies, and we surprisingly found it did not cause much power loss in our GWA scan for plasma Lp(a) levels. The underlying reason is that the correlation value between ancestry and a single SNP is low, even for an ancestry-informative one, which seems justifying usage of this approach in GWA studies. However, there is no guarantee this correlation is always low in other admixed populations, and population stratification is not the only resource of confounding in genetic studies; moreover, the conclusions this study draws are not limited to confounding factors, they apply to all risk/protective factors. Therefore, caution still needs to be taken when employing this approach in genetic studies. In summary, we advocate using approaches I and II instead of approaches III or IV in practice when adjusting for potential confounding factors.

Acknowledgments

We thank the reviewers' constructive comments which tremendously improved the manuscript from an earlier version, and we thank Dr. Helen Hobbs for granting us permission to use the DHS data. This study is supported by the American Heart Association Scientist Development Grant 10SDG4220051 to C.X.

Appendix

Least Squares Estimate $\hat{\gamma}_i$ of γ_i

Consider a linear regression model $Z = M\lambda + \varepsilon$ (g), where Z is the dependent variable, M is the design matrix with the first column to be $J = (1, 1, \dots, 1)'$ to incorporate an intercept term in the model, and ε is the random error term with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma_Z^2$. A general formula to calculate the least squares estimators of regression coefficients is $\hat{\lambda} = (M'M)^{-1} M'Z$. Based on this formula, below we calculate $\hat{\gamma}_i$, where $i \in \{I, II, III, IV\}$, for each of the 4 approaches.

Suppose X is an $n \times p$ matrix with the first column to be $J = (1, 1, \dots, 1)'$ to incorporate an intercept term in the model, i.e. there are n subjects in the sample and $p - 1$ covariates to be adjusted for. For approach I, the design matrix in model a is $M_I = [X, G]$, and thus the least squares estimate of γ_I can be computed as $\hat{\gamma}_I = [(M_I'M_I)^{-1} M_I'Y]_{p+1,1}$. Alternatively, we can calculate the estimate by

$$\hat{\gamma}_I = \frac{G'R_xY}{G'R_xG},$$

where $R_x = I_n - X(X'X)^{-1} X'$ [23]. For approach II, the design matrix in model d is $M_{II} = [J, R_xG]$. Note that $J'R_xG = 0$ and $R_x'R_x = R_x$, thus the least squares estimate

$$\hat{\gamma}_{II} = \left[(M_{II}'M_{II})^{-1} M_{II}'R_xY \right]_{2,1} = \frac{G'R_xY}{G'R_xG},$$

which is equal to $\hat{\gamma}_I$. For approach III, the design matrix in model e is $M_{III} = [J, G]$. Note that $J'R_xY = 0$, thus the least squares estimate

$$\hat{\gamma}_{III} = \left[(M_{III}'M_{III})^{-1} M_{III}'R_xY \right]_{2,1} = \frac{G'R_xY}{G'G - (\sum G)^2/n}.$$

For approach IV, the design matrix in model f is the same as that in model d, i.e. $M_{IV} = M_{II} = [J, R_xG]$. Thus the least squares estimate

$$\hat{\gamma}_{IV} = \left[(M_{IV}'M_{IV})^{-1} M_{IV}'Y \right]_{2,1} = \frac{G'R_xY}{G'R_xG},$$

which is equal to $\hat{\gamma}_I$ and $\hat{\gamma}_{II}$.

The difference between $\hat{\gamma}_{III}$ and the other three estimates is their denominators, which are $G'G - (\sum G)^2/n$ and $G'R_xG$, respectively. Note that for the regression model c, $SSTO = G'(I_n - JJ'/n)G = G'G - (\sum G)^2/n$, and $SSE = G'R_xG$. Therefore, there exists an inequality $|\hat{\gamma}_{III}| < |\hat{\gamma}_I = \hat{\gamma}_{II} = \hat{\gamma}_{IV}|$ because $SSTO > SSE$. Furthermore, the relative bias

$$\frac{\hat{\gamma}_I - \hat{\gamma}_{III}}{\hat{\gamma}_I} = \frac{SSTO - SSE}{SSTO} = R^2,$$

which is the coefficient of determination of model c.

Variance Estimate $Var(\hat{\gamma}_i)$

Again consider the linear regression model g. A general formula to calculate the variance of regression coefficients is $Var(\lambda) = \sigma_Z^2(M'M)^{-1}$, in which σ_Z^2 is ordinarily estimated by the error mean square

$$MSE = \frac{SSE}{n - q},$$

where q is the number of columns of the design matrix. Based on these formulae, below we calculate $Var(\hat{\gamma}_i)$, for each of the 4 approaches.

For approach I, $Var(\hat{\gamma}_I) = \sigma_I^2[(M_I'M_I)^{-1}]_{p+1,p+1}$. Alternatively, given

$$\hat{\gamma}_I = \frac{G'R_xY}{G'R_xG},$$

it can be shown

$$Var(\hat{\gamma}_I) = \frac{\sigma_I^2}{G'R_xG}.$$

Define $R_I = I_n - M_I(M_I'M_I)^{-1} M_I'$. The variance of Y is estimated by

$$\hat{\sigma}_I^2 = \frac{(R_IY)'R_IY}{n - p - 1} = \frac{Y'R_IY}{n - p - 1}.$$

It can be further shown that

$$\hat{\sigma}_I^2 = \frac{Y'R_xY - \hat{\gamma}_I G'R_xY}{n - p - 1} = \frac{Y'R_xY - \frac{(G'R_xY)^2}{G'R_xG}}{n - p - 1}$$

[23]. Therefore

$$Var(\hat{\gamma}_I) = \frac{Y'R_xY - \frac{(G'R_xY)^2}{G'R_xG}}{(n - p - 1)G'R_xG}.$$

For approach II,

$$Var(\hat{\gamma}_{II}) = \sigma_{II}^2 \left[(M_{II}'M_{II})^{-1} \right]_{2,2} = \frac{\sigma_{II}^2}{G'R_xG},$$

and

$$\hat{\sigma}_{II}^2 = \frac{(R_{II}R_xY)'R_{II}R_xY}{n - 2},$$

where $R_{II} = I_n - M_{II}(M_{II}'M_{II})^{-1} M_{II}'$. After some matrix operations, it can be shown that

$$\hat{\sigma}_{II}^2 = \frac{Y'R_xY - \frac{(G'R_xY)^2}{G'R_xG}}{n - 2},$$

and thus

$$Var(\hat{\gamma}_{II}) = \frac{Y'R_xY - \frac{(G'R_xY)^2}{G'R_xG}}{(n - 2)G'R_xG}.$$

For approach III,

$$Var(\hat{\gamma}_{III}) = \sigma_{III}^2 \left[(M_{III}'M_{III})^{-1} \right]_{2,2} = \frac{\sigma_{III}^2}{G'G - (\sum G)^2/n},$$

and

$$\hat{\sigma}_{III}^2 = \frac{(R_{III}R_xY)'R_{III}R_xY}{n - 2},$$

where $R_{III} = I_n - M_{III}(M_{III}'M_{III})^{-1} M_{III}'$. After some matrix operations, it can be shown that

$$\hat{\sigma}_{III}^2 = \frac{Y'R_xY - \frac{(G'R_xY)^2}{G'G - (\sum G)^2/n}}{n - 2},$$

and therefore

$$\text{Var}(\hat{\gamma}_{III}) = \frac{Y'R_xY - \frac{(G'R_xY)^2}{G'G - (\sum G)^2/n}}{(n-2)[G'G - (\sum G)^2/n]}.$$

For approach IV,

$$\text{Var}(\hat{\gamma}_{IV}) = \sigma_{IV}^2 [(M_{II}'M_{II})^{-1}]_{2,2} = \frac{\sigma_{IV}^2}{G'R_xG},$$

and

$$\hat{\sigma}_{IV}^2 = \frac{(R_{II}Y)'R_{II}Y}{n-2}.$$

It can be further shown that

$$\hat{\sigma}_{IV}^2 = \frac{Y'Y - (\sum Y)^2/n - \frac{(G'R_xY)^2}{G'R_xG}}{n-2},$$

and thus

$$\text{Var}(\hat{\gamma}_{IV}) = \frac{Y'Y - (\sum Y)^2/n - \frac{(G'R_xY)^2}{G'R_xG}}{(n-2)G'R_xG}.$$

Note that the difference between $\text{Var}(\hat{\gamma}_I)$, $\text{Var}(\hat{\gamma}_{II})$, and $\text{Var}(\hat{\gamma}_{IV})$ lies in their estimates of σ_i^2 . Models a and d have the same SSE but different degrees of freedom: $n-p-1$ and $n-2$, respectively. Therefore

$$\frac{\text{Var}(\hat{\gamma}_I)}{\text{Var}(\hat{\gamma}_{II})} = \frac{n-2}{n-p-1}$$

and

$$\lim_{n \rightarrow \infty} \frac{\text{Var}(\hat{\gamma}_I)}{\text{Var}(\hat{\gamma}_{II})} = 1.$$

Models d and f have different SSE in spite of the same degrees of freedom. Note that for regression model b $SSTO = Y'Y - (\sum Y)^2/n$ and $SSE = Y'R_xY$. Hence we have $\text{Var}(\hat{\gamma}_{IV}) > \text{Var}(\hat{\gamma}_{II})$.

$$\text{Wald's Test Statistic } W_i = \hat{\gamma}_i^2 / \text{Var}(\hat{\gamma}_i)$$

Given $\hat{\gamma}_I = \hat{\gamma}_{II}$ and

$$\frac{\text{Var}(\hat{\gamma}_I)}{\text{Var}(\hat{\gamma}_{II})} = \frac{n-2}{n-p-1},$$

it is straightforward that

$$\frac{W_I}{W_{II}} = \frac{n-p-1}{n-2}$$

and

$$\lim_{n \rightarrow \infty} \frac{W_I}{W_{II}} = 1.$$

Given $\hat{\gamma}_{II} = \hat{\gamma}_{IV}$ and $\text{Var}(\hat{\gamma}_{IV}) > \text{Var}(\hat{\gamma}_{II})$, it is straightforward that $W_{II} > W_{IV}$. It can be shown that

$$W_{II} = \frac{\hat{\gamma}_{II}^2}{\text{Var}(\hat{\gamma}_{II})} = \frac{(n-2)(G'R_xY)^2}{(G'R_xG)(Y'R_xY) - (G'R_xY)^2}$$

and

$$W_{III} = \frac{\hat{\gamma}_{III}^2}{\text{Var}(\hat{\gamma}_{III})} = \frac{(n-2)(G'R_xY)^2}{[G'G - (\sum G)^2/n](Y'R_xY) - (G'R_xY)^2}.$$

Note that $G'G - (\sum G)^2/n$ and $G'R_xG$ are SSTO and SSE, respectively, for regression model c. Therefore, we conclude $W_{II} > W_{III}$.

References

- Rothman KJ, Greenland S, Lash TL: Modern Epidemiology, ed 3. Philadelphia, Lippincott Williams & Wilkins, 2008.
- Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70:41–55.
- Rosenbaum PR, Rubin DB: Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984;79:516–524.
- Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52:506–516.
- Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999;55:997–1004.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. *Am J Hum Genet* 2000; 67:170–181.
- 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;38:904–909.
- Satten GA, Flanders WD, Yang Q: Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001;68:466–477.
- Allen AS, Satten GA: Control for confounding in case-control studies using the stratification score, a retrospective balancing score. *Am J Epidemiol* 2011;173:752–760.
- Epstein MP, Allen AS, Satten GA: A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet* 2007;80:921–930.
- Zhao H, Rebbeck TR, Mitra N: A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors. *Genet Epidemiol* 2009; 33:679–690.
- Xing C, Cohen JC, Boerwinkle E: A weighted false discovery rate control procedure reveals alleles at FOXA2 that influence fasting glucose levels. *Am J Hum Genet* 2010;86: 440–446.
- Deo RC, Reich D, Tandon A, Akyzbekova E, Patterson N, Waliszewska A, Kathiresan S, Sarpong D, Taylor HA Jr, Wilson JG: Genetic differences between the determinants of lipid profile phenotypes in African and European Americans: the Jackson Heart Study. *PLoS Genet* 2009;5:e1000342.
- Paschou P, Drineas P, Lewis J, Nievergelt CM, Nickerson DA, Smith JD, Ridker PM, Chasman DI, Krauss RM, Ziv E: Tracing substructure in the European American population with PCA-informative markers. *PLoS Genet* 2008;4:e1000114.
- The 1000 Genomes Project Consortium: A map of human genome variation from population-scale sequencing. *Nature* 2010;467: 1061–1073.

- 16 Su Z, Marchini J, Donnelly P: Hapgen2: simulation of multiple disease SNPs. *Bioinformatics* 2011;27:2304–2305.
- 17 Guyton JR, Dahlen GH, Patsch W, Kautz JA, Gotto AM Jr: Relationship of plasma lipoprotein lp(a) levels to race and to apolipoprotein b. *Arteriosclerosis* 1985;5:265–272.
- 18 Boerwinkle E, Leffert CC, Lin J, Lackner C, Chiesa G, Hobbs HH: Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. *J Clin Invest* 1992;90:52–60.
- 19 Boerwinkle E, Menzel HJ, Kraft HG, Utermann G: Genetics of the quantitative lp(a) lipoprotein trait. III. Contribution of lp(a) glycoprotein phenotypes to normal lipid variation. *Hum Genet* 1989;82:73–78.
- 20 Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA, Staab JM, Hobbs HH: The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. *Am J Cardiol* 2004;93:1473–1480.
- 21 Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, Malasky MJ, Scafe C, Le E, De Jager PL, Mignault AA, Yi Z, De The G, Essex M, Sankale JL, Moore JH, Poku K, Phair JP, Goedert JJ, Vlahov D, Williams SM, Tishkoff SA, Winkler CA, De La Vega FM, Woodage T, Sninsky JJ, Hafler DA, Altshuler D, Gilbert DA, O'Brien SJ, Reich D: A high-density admixture map for disease gene discovery in African Americans. *Am J Hum Genet* 2004;74:1001–1013.
- 22 Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, Boerwinkle E, Cohen JC, Hobbs HH: Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* 2008;40:1461–1465.
- 23 Seber GAF, Lee AJ: *Linear Regression Analysis*, ed 2. Hoboken, John Wiley & Sons, Inc., 2003.
- 24 Rosenbaum PR, Rubin DB: Difficulties with regression analyses of age-adjusted rates. *Biometrics* 1984;40:437–443.
- 25 Cox DR, Hinkley DV: *Theoretical Statistics*, ed 1. Chapman & Hall/CRC, 1974.
- 26 Vandaele W: Wald, likelihood ratio, and Lagrange multiplier tests as an F test. *Economics Letters* 1981;8:361–365.
- 27 Neuhaus JM: Estimation efficiency with omitted covariates in generalized linear models. *J Am Stat Assoc* 1998;93:1124–1129.
- 28 Xing G, Xing C: Adjusting for covariates in logistic regression models. *Genet Epidemiol* 2010;34:769–771; author reply 772.