Published in final edited form as:
Genet Epidemiol. 2010 November ; 34(7): 769–772. doi:10.1002/gepi.20526.

Adjusting for Covariates in Logistic Regression Models

Guan Xing[1] and **Chao Xing**[2,3]

[1]Bristol-Myers Squibb Company, Pennington, New Jersey

[2]Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas

[3]McDermott Center of Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas

To the Editor

We read with interest the paper by Kuo and Feingold (henceforth K&F) "What's the best statistic for a simple test of genetic association in a case-control study" [Kuo and Feingold, 2010], in which the authors compared the power of three logistic regression models to detect genetic effects, concluded that "the most commonly used approach to handle covariates—modeling covariate main effects but not interaction—is almost never a good idea," and recommended modeling only the genetic factors without covariate adjustment in genome-scanning. We feel that the issue of covariate adjustment in logistic regression models was oversimplified and the conclusion was unjustified. In this letter, we attempt to explain the observations found by K&F using established results in the statistical literature, to confirm the theoretical results in the genetic association study scenario by mimicking and extending the simulation study performed by K&F, and to discuss their implication on covariate adjustment in genome-scanning.

The impact of covariate adjustment on the precision of regression coefficient estimators in classic linear models depends on multiple correlations between variables; however, adjusting for covariates in logistic regression models always leads to a loss of precision. Denote by Y a quantitative trait, by G a genotype, and by E a covariate, e.g. an environmental factor. Suppose we fit to data two linear regression models $E(Y|G) = \alpha+\beta_1 G$ and $E(Y|G, E) = \alpha' + \beta'_1 G + \beta'_2 E$. The asymptotic relative precision (ARP) of the estimator $\hat{\beta}'_1$ to $\hat{\beta}_1$ can be derived as

$$ARP(\widehat{\beta}'_1 \text{ to } \widehat{\beta}_1)=\text{Var}(\widehat{\beta}_1)/\text{Var}(\widehat{\beta}'_1)=(1-\rho^2_{GE})/(1-\rho^2_{YE|G}) \quad (1)$$

where ρ_{GE} is the correlation coefficient between G and E, and $\rho_{YE|G}$ is the partial correlation coefficient between Y and E given G. Thus, whether ARP ($\hat{\beta}'_1$ to $\hat{\beta}_1$) is greater or smaller than one depends on both the correlations—between G and E, and between Y and E [Robinson and Jewell, 1991]. On the other hand, adjusting for covariates in logistic regression models leads to increased variances of coefficient estimators regardless of correlations between variables. Denote by $D\varepsilon\{0,1\}$ affection status. Suppose we fit to data the two logistic regression models (a): logit $P(D = 1|G) = \alpha+\beta_1 G$ and (b): logit $P(D = 1|G, E) = \alpha' + \beta'_1 G + \beta'_2 E$. It can be shown that ARP ($\hat{\beta}'_1$ to $\hat{\beta}_1 \leq 1$, i.e. formula (1), established in classic linear regression models, breaks down in logistic regression models; moreover, the larger the magnitude of $\hat{\beta}'_2$, the poorer the precision of the estimator $\hat{\beta}'_1$ [Wickramaratne and Holford, 1990; Robinson and Jewell, 1991]. The increase in variance of $\hat{\beta}'_1$ over that of $\hat{\beta}_1$ can lead to a power loss in testing the null hypothesis of no genetic effect when $\hat{\beta}'_1$ is

asymptotically equal to $\hat{\beta}_1$, which explains the results of model 1—genetic effect only—in K&F, i.e. model (*a*) is more powerful than model (*b*).

Now when model (*b*) underlies the disease susceptibility and *G* and *E* are independent, omitting the environmental factor leads to a downward bias in estimating effect sizes of the genetic factor, i.e., $|\hat{\beta}_1| < |\hat{\beta}'_1|$ [Gail et al., 1984; Neuhaus and Jewell, 1993]. Noting that Var ($\hat{\beta}_1$) < Var ($\hat{\beta}'_1$), it is of interest to investigate which model, (a) or (b), is more powerful in testing the null hypothesis of no genetic effect. It has been shown that the asymptotic relative efficiency (ARE) of the two hypothesis tests meets the inequality ARE ($\hat{\beta}'_1$ to $\hat{\beta}_1$ at $\beta_1 = 0$) ≥ 1 [Robinson and Jewell, 1991; Neuhaus, 1998]. Therefore, in testing the null hypothesis of no genetic effect, omitting disease risk/preventive factors that are independent of genetic factors always leads to efficiency loss, which explains the results of model 2—genetic and environmental marginal effects only—in K&F, i.e. model (b) is more powerful than model (a). Note that K&F did not clearly state which model was more powerful, though their figures showed it was model (b), which we confirmed by mimicking their simulation study as presented below. Furthermore, the magnitude of the loss increases with the strength of the association between the omitted covariate and the disease [Neuhaus, 1998].

We first simulated case-control data with a single-locus disease model and a covariate using the method by K&F. For simplicity, we only simulated an additive genetic model with the penetrance functions summarized in Table I, in which models 1, 2a, 3a, and 3b are identical to the models 1, 2, 3, and 4, respectively, in Supplemental Table 1 of K&F. Model 1 contains a genetic effect only. Models 2a–2c have genetic and environmental marginal effects only. From models 2a to 2c, the effect size of the environmental factor increases. Models 3a and 3b include both marginal and interaction effects of genetic and environmental factors. From models 3a to 3b, the interaction effect size decreases. In all the models, we set the minor allele frequency equal to 0.3, probability of exposure equal to 0.5, and the genetic and environmental factors are independent in the general population. Under each model, we simulated 10,000 replicates with 300 cases and 300 controls in each data set. We analyzed each data set by both the models (a) and (b), and compared the coefficient estimates, their standard errors, and power to detect the genetic effect (Table II). As predicted by ARP ($\hat{\beta}'_1$ to $\hat{\beta}_1$) ≤ 1 in logistic regression models, the standard error of $\hat{\beta}_1$ was smaller than that of $\hat{\beta}'_1$ in all the six genetic models. For model 1, although there was no statistical difference between the estimates $\hat{\beta}'_1$ and $\hat{\beta}_1$, the analysis model (*b*) was slightly less powerful than model (a) owing to a decreased precision of $\hat{\beta}'_1$ compared with $\hat{\beta}_1$. For models 2a–2c, both the magnitude and standard error of $\hat{\beta}_1$ were smaller than those of $\hat{\beta}'_1$, and the power of model (b) was greater than that of model (a) in testing for the genetic effect; all the differences became more significant as the environmental effects increased from models 2a to 2c. There has been no theoretical study investigating the estimation bias and efficiency with omitted covariates and interaction terms when the underlying true model includes both marginal and interaction effects, and we speculate the results depend on the relative magnitude and direction of both effects. In models 3a and 3b, the marginal and interaction effects were in the same direction, and the magnitude of $\hat{\beta}_1$ was greater than that of $\hat{\beta}'_1$. An intuitive explanation is that a larger proportion of the interaction effect is absorbed into the estimated marginal effect $\hat{\beta}_1$ than into $\hat{\beta}'_1$ resulting in $\hat{\beta}_1 > \hat{\beta}'_1$. The greater coefficient estimate, together with a smaller standard error, led to the greater power of model (a) than model (b) in testing the genetic effect.

We then carried out simulations under the logistic regression model to examine the magnitude of power loss due to omitting predictive covariates that are independent of the genetic effect. Assuming the true disease model was (b), in which α' was chosen such that the disease prevalence equaled to 0.15, and β'_1 was chosen such that the odds ratio (OR) of per copy of disease-disposing allele equaled to 1.2, we simulated a diallelic marker with the

minor allele frequency equal to 0.3, and a binary covariate independent of the genetic variant with the OR $\varepsilon\{1.5,2.0,2.5,3.0,3.5,4.0,4.5,5.0\}$ and the exposure frequency $\varepsilon\{0.2,0.5\}$. The parameter settings mimicked those of a typical complex trait suggested by results of genome-wide association studies. Under each scenario 10,000 replicates with 1,000 cases and 1,000 controls in each data set were simulated. We analyzed each data set by both models (a) and (b), and compared their power to detect the genetic effect (Table III). Omitting the covariate always led to power loss, and the magnitude of the loss increased with the effect size of the covariate. When the OR was less than 3.0, the power loss was trivial; however, the loss can be substantial when the OR increased to 5.0. For example, if there were disparity of disease prevalence between males and females, omitting the covariate sex would result in considerable power loss.

A conventional wisdom in classic linear regression is that adjusting for covariates associated with the response variable can improve the precision of estimates by reducing the residual variance [Fisher, 1932]; however, covariate adjustment in logistic regression models always leads to a loss of precision. Nonetheless, this loss of precision does not always result in a loss of power. When the genetic and risk/preventive environmental factors are independent and do not have interaction effects on the disease, it is always more efficient to adjust for the predictive covariates. Note that this conclusion holds not only in logistic regression models but also in the class of generalized linear models [Neuhaus, 1998]. A person's genetic background is determined from birth, thus in most cases it is not unreasonable to assume it is independent of his/her subsequent environmental exposure, which has been a key assumption in some study designs to investigate gene-environmental interaction [Piegorsch et al., 1994; Chatterjee and Carroll, 2005]. If we assume that only a small proportion of genetic variants interact with the known environmental factors on disease susceptibility, then, contrary to the conclusion by K&F, we recommend adjusting for predictive covariates at the genome-scanning stage.

Acknowledgments

We thank Dr. Robert Elston for critical reading of the manuscript and helpful discussions. CX was partially supported by a Pilot Award from UL1RR024982 from the National Center for Research Resources.

References

Chatterjee N, Carroll RJ. Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. Biometrika. 2005; 92:399–418.

Fisher, RA. Statistical Methods for Research Workers. 14th. Edinburgh: Oliver & Boyd; 1970.

Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984; 71:431–444.

Kuo CL, Feingold E. What's the best statistic for a simple test of genetic association in a case-control study? Genet Epidemiol. 2010; 34:246–253. [PubMed: 20025064]

Neuhaus JM. Estimation efficiency with omitted covariates in generalized linear models. J Am Stat Assoc. 1998; 93:1124–1129.

Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. Biometrika. 1993; 80:807–815.

Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med. 1994; 13:153–162. [PubMed: 8122051]

Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. Int Stat Rev. 1991; 58:227–240.

Wickramaratne PJ, Holford TR. Confounders: correcting superstitions. Authors' reply. Biometrics. 1990; 46:870–872.

TABLE I

Penetrance functions of additive genetic models[a]

	Exposed			Non-exposed		
Disease model	f_0	f_1	f_2	f_0	f_1	f_2
1	0.01	0.015	0.02	0.01	0.015	0.02
2a	0.015	0.02	0.025	0.01	0.015	0.02
2b	0.02	0.025	0.03	0.01	0.015	0.02
2c	0.025	0.03	0.035	0.01	0.015	0.02
3a	0.01	0.015	0.02	0.01	0.01	0.01
3b	0.01	0.02	0.03	0.01	0.015	0.02

[a]Minor allele frequency equals 0.3; probability of exposure equals 0.5; f_i denotes the probability of being affected given $i \varepsilon \{0, 1, 2\}$ number of copies of the minor allele.

TABLE II

Comparison of coefficient estimates, their standard errors (SE), and power of logistic regression models with and without covariate adjustment[a]

	Comparison of $\hat{\beta}_1$ and $\hat{\beta}'_1$		Comparison of SE($\hat{\beta}_1$) and SE($\hat{\beta}'_1$)		
Disease model	**Percentage of $\lvert\hat{\beta}_1\rvert < \lvert\hat{\beta}'_1\rvert$**	***P*-value[b]**	**Percentage of SE($\hat{\beta}_1$) < SE($\hat{\beta}'_1$)**	***P*-value[b]**	**Power difference[c]**
1	0.5391	7.0×10^{-1}	0.9828	4.2×10^{-16}	0.0003
2a	0.6567	7.0×10^{-3}	0.9981	6.7×10^{-54}	−0.0101
2b	0.7245	8.7×10^{-11}	1.0000	3.8×10^{-285}	−0.0222
2c	0.7554	1.4×10^{-24}	1.0000	$<3.8 \times 10^{-285}$	−0.0441
3a	0.1630	5.2×10^{-10}	0.9991	3.5×10^{-50}	0.0399
3b	0.2792	1.2×10^{-2}	0.9904	1.2×10^{-16}	0.0019

[a] The results are based on 10,000 replicates of 300 cases and 300 controls. The two logistic models are (a): logit $P(D = 1|G) = \alpha + \beta_1 G$ and (b): logit $P(D = 1|G, E) = \alpha' + \beta'_1 G + \beta'_2 E$.

[b] Two-sample *t*-test to test the null hypotheses of $|\hat{\beta}_1| = |\hat{\beta}'_1|$ and $SE(\hat{\beta}_1) = SE(\hat{\beta}'_1)$.

[c] Empirical power of model (a) minus that of model (b) at the significance level of 0.05.

TABLE III

Impact of omitting a predictive covariate on the power to detect the genetic effect by logistic regression models[a]

	Exposure frequency	
Odds ratio	**0.2**	**0.5**
1.5	0.0005	0.0007
2.0	0.0028	0.0072
2.5	0.0084	0.0083
3.0	0.0155	0.0146
3.5	0.0189	0.0206
4.0	0.0215	0.0282
4.5	0.0332	0.0381
5.0	0.0419	0.0476

[a]The empirical power of model (b): logit $P(D = 1|G, E) = \alpha' + \beta'_1 G + \beta'_2 E$ minus that of model (*a*): logit $P(D = 1|G) = \alpha + \beta_1 G$ at the significance level of 0.05 is reported. The results are based on 10,000 replicates of 1,000 cases and 1,000 controls. The disease prevalence equals 0.15, the frequency of the disease-disposing allele equals 0.3, and its odds ratio equals 1.2.