Published in final edited form as:

Epidemiology. 2011 March ; 22(2): 257–261. doi:10.1097/EDE.0b013e31820877c5.

On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is mis-specified

Eric J. Tchetgen Tchetgen and Peter Kraft

Departments of Epidemiology and Biostatistics, Harvard University

Abstract

We consider the robustness of tests of genetic associations that incorporate gene-environment interactions when the environmental exposure is mis-specified, which is likely the case when the exposure is continuous. We note that under the null hypothesis of no genetic association, mis-specified ordinary logistic regression and profile likelihood analyses both consistently estimate the null parameters of no genetic main effect and interaction, provided that genetic and environmental factors are unrelated in the underlying population. However, we argue that the associated standard likelihood ratio test statistic and other related test statistics have incorrect type 1 error rate due to model mis-specification. Based on these observations, we propose the use of the sandwich estimator of variance in conjunction with the consistent maximum (profile) likelihood estimates to construct Wald-type tests statistics with correct type 1 error rate for the null of no genetic association.

1 Robustness of logistic regression

Standard prospective logistic regression analysis is the most commonly used analytic approach to evaluate genetic associations from case-control data in the presence of gene-environment interaction (Kraft et al, 2007). In the simple setting where the genetic variant G and the environmental factor E are both binary, the approach entails fitting by maximum likelihood a saturated ordinary logistic regression of disease status D , where $D = 1$ for cases and $D = 0$ for controls:

$$\text{logit Pr}\{D = 1 | G, E; \beta\} = \beta_0 + \beta_E E + \beta_G G + \beta_{GE} G \times E \quad (1)$$

Here $\beta = (\beta_0, \beta_G, \beta_E, \beta_{GE})$ represents the unknown regression parameters. A test of the null hypothesis of no genetic associations $\beta_G = \beta_{GE} = 0$ can then be obtained by using a standard 2 degree of freedom likelihood ratio chi-squared statistic, or alternatively by using its Wald or score statistic analogues. Similarly, a 1 degree of freedom test of no effect measure modification $\beta_{GE} = 0$, i.e. no departure from a multiplicative odds model for the joint effect of G and E , can be implemented. In practice, the environmental exposure measured in a case-control study is often either of a polytomous or of a continuous nature. For example, a measurement of body mass index classified as one of four possible categories (underweight, normal, overweight and obese) would constitute a polytomous environmental exposure,

whereas body mass index recorded on its original scale (mass in kilograms/(height in meters)²) constitutes a continuous environmental exposure. Hereafter, for simplicity, we only consider the continuous case although the polytomous case is easily deduced from the results.

In the presence of a continuous exposure variable, the regression model given in equation (1) may no longer be correctly specified, because the model assumes that the logarithm of the odds ratio association between E and D within levels of G is linear in E . This assumption may be erroneous in practice. For example, the observed data may have been generated from a model in which the true relationship between the exposure and log odds of disease among controls followed a quadratic form $\beta_E E + \beta_{E^2} E^2$ rather than the linear form $\beta_E E$ of equation (1). The extent to which this form of model mis-specification can have a detrimental impact on inferences on genetic effects has been of recent interest in the genetic epidemiology literature (Cornelis et al. 2010, Tchetgen Tchetgen, 2010).

For example, in a specific application to a Genome-wide interaction study of type 2 diabetes, Cornelis et al (2010) noted that a mis-specified main effect of continuous body mass index leads to a dramatic inflation of the genomewide type 1 error rate of Wald test statistics of the null hypothesis of no genetic associations (main effect and/or interaction with body mass index) over all candidate genetic variants. They also noted that, not surprisingly, upon correcting the model mis-specification by accounting for the presence of a nonlinear effect of body mass index, the type 1 error rate of the corresponding tests returned to its nominal level. Interestingly, they further noted that in an additional analysis of their initial mis-specified models, a substitution of Huber's sandwich estimator for the inverse information matrix estimator of covariance in the construction of Wald-type statistics recovered a type 1 error rate at the nominal level. This latter observation suggests a level of robustness of ordinary logistic regression.

In the current paper, we formally show that under the null hypothesis of no genetic effects, ordinary logistic regression will correctly estimate null genetic associations, even if the model for the odds ratio association **or** ($E|G=0$) relating the continuous exposure variable to the disease outcome among non-carriers is incorrectly specified, provided that the genetic and environmental factors are unrelated in the underlying population. This remarkable fact implies that although ordinary logistic regression does not exploit the assumption of gene-environment independence to improve estimation efficiency, it can exploit the independence assumption to construct tests of genetic effects with guaranteed nominal type 1 error level. Next, in order to provide intuition for this result, we give an argument as to its correctness in the simple case of the working model given in equation (1) where E is now continuous. For a general formalization of the result, we refer the reader to Rosenblum and van der Laan (2008) who provide a comprehensive study of robustness under the null hypothesis in the context of a prospective randomized experiment. Although of a retrospective nature, the current setting is analogous to a randomized prospective experimental study, to the extent that under the null hypothesis and the independence assumption, the genetic variant behaves, for all practical purposes, similar to a randomized exposure. Their result also shows that a similar property holds for cohort data, with polytomous, count or continuous outcome,

randomized polytomous exposure, and a variety of link functions including the canonical link functions associated with densities in the exponential family.

To proceed, suppose that $\log \text{or}(E|G=0)$ is in reality a nonlinear function of E in the underlying population of interest. Then, consider fitting by maximum likelihood the working logistic regression model

$$\text{logit } \Pr^*(D = 1|G, E; \theta) = \theta_1 + \theta_2 G + \theta_3 GE + \theta_4 E \equiv \log\{\text{odds}_{G,E}(G, E; \theta)\} \quad (2)$$

Here $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, and the '*' indicates that the regression is now a working model which is not necessarily assumed to be correct. Despite of model (2) being mis-specified, the result states that under the null hypothesis of no genetic effects, the maximum likelihood estimator $(\hat{\theta}_2, \hat{\theta}_3)$ correctly converges (in probability) to $(\beta_G, \beta_{GE}) = (0, 0)$. The estimate $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)$ solves the following logistic regression score equation

$$\sum_i [1, G_i, G_i E_i, E_i]^T \{D_i - \Pr^*(D = 1|G, E; \hat{\theta})\} = 0 \quad (3)$$

To see why $(\hat{\theta}_2, \hat{\theta}_3)$ is consistent for $(0, 0)$, let $\Pr(G_i = 1)$ denote the proportion of individuals in the underlying population with $G = 1$. The score equation (3) implies that

$$\sum_i [1, E_i]^T (G_i - \Pr(G_i = 1)) \{D_i - \Pr^*(D = 1|G, E; \hat{\theta})\} = 0$$

which is obtained by multiplying the first and forth equation in (3) by $\Pr(G_i = 1)$, then subtracting them from the second and third equations, respectively. Under standard regularity conditions, $\hat{\theta}$ converges (in probability) to the value θ^* of θ which satisfies the "population" version of this last equation:

$$0 = E\{[1, E]^T (G - \Pr(G = 1)) (D - \Pr^*(D = 1|G, E; \theta^*))\}$$

The expectation is taken over $[E, G, D]$ for cross-sectional or prospectively-collected data, or over $[E, G|D]$ for case-control data. This equation can be simplified as follows:

$$\begin{aligned} 0 &= E\{[1, E]^T (G - \Pr(G = 1)) D\} - E\{[1, E]^T (G - \Pr(G = 1)) \Pr^*(D = 1|G, E; \theta^*)\} \quad (4) \\ &= -E\{[1, E]^T (G - \Pr(G = 1)) \Pr^*(D = 1|G, E; \theta^*)\} \\ &= -\text{Var}(G) \times E\{[1, E]^T [\Pr^*(D = 1|G = 0, E; \theta^*) - \Pr^*(D = 1|G = 1, E; \theta^*)]\} \end{aligned}$$

Here we have used the null hypothesis and the gene -environment independence assumption which together imply that

$$E\{[1, E]^T (G - \Pr(G = 1)) D\} = E\{[1, E]^T E[(G - \Pr(G = 1)) | D, E] D\} = E\{[1, E]^T E[(G - \Pr(G = 1)) | D] D\} = 0$$

The expression in equation (4) is zero if and only if

$$\Pr^*(D = 1|G = 0, E; \theta^*) - \Pr^*(D = 1|G = 1, E; \theta^*) = 0,$$

therefore θ_2^* and θ_3^* must both be equal to zero. Thus, we have established that in spite of model mis-specification, the ordinary logistic regression maximum likelihood estimator of (θ_2, θ_3) converges correctly to $(\beta_G, \beta_{GE}) = (0, 0)$, provided that G and E are independent in the underlying population.

However, in spite of the unbiasedness (consistency) of the maximum likelihood estimator $(\hat{\theta}_2, \hat{\theta}_3)$, the standard likelihood ratio statistic and the Wald statistic to test the null hypothesis of no genetic effects do not follow the usual chi-squared distribution because the usual inverse of the information matrix estimator of their covariance matrix is expected to be biased even under the null. The large sample covariance matrix of $\hat{\theta}$ is given by

$$\{E(\dot{S}_i(\theta^*))\}^{-1} E\{S_i(\theta^*)S_i(\theta^*)^T\} \{E(\dot{S}_i(\theta^*))\}^{-1} \quad (5)$$

where $S_i(\theta)$ and $\dot{S}_i(\theta)$ are respectively the i th individual's contribution to the score equation and Hessian matrix with respect to θ under the working model (2), evaluated at θ . In the standard setting where $\Pr^*(D = 1|G, E; \theta^*)$ is identical to the true data generating mechanism $\Pr(D = 1|G, E)$,

$$E\{S_i(\theta^*)S_i(\theta^*)^T\} = -E(\dot{S}_i(\theta^*))$$

by the well-known information equality, leading to simplification of equation (5) to the familiar inverse-information formula of the covariance matrix. In the current setting, $\Pr^*(D = 1|G, E; \theta^*)$ does not equal $\Pr(D = 1|G, E)$, thus the above simplification does not hold. Because of this, tests using the standard inverse-information estimate for the covariance of $\hat{\theta}$ will have the incorrect size, and can be inflated as in the case of Cornelis et al (2010), or deflated.

We recommend that the matrix $\hat{\Sigma}$ be used to estimate the covariance of $(\hat{\theta}_2, \hat{\theta}_3)$, where $\hat{\Sigma}$ is obtained by extracting the sub-matrix for the parameters (θ_2, θ_3) from Huber's sandwich estimator of the covariance matrix (Huber, 1967):

$$\left\{ \sum_i \dot{S}_i(\hat{\theta}) \right\}^{-1} \sum_i S_i(\hat{\theta}) S_i(\hat{\theta})^T \left\{ \sum_i \dot{S}_i(\hat{\theta}) \right\}^{-1} \quad (6)$$

which is an empirical version of equation (5). Then, to formally test the null hypothesis of no genetic effects, we recommend that the analyst use the following Wald-type test statistics constructed with $(\hat{\theta}_2, \hat{\theta}_3)$ and $\hat{\Sigma}$:

$$[\hat{\theta}_2, \hat{\theta}_3] \widehat{\Sigma}^{-1} [\hat{\theta}_2, \hat{\theta}_3]^T \sim \chi^2_2 \text{ under the null of no genetic effects} \quad (7)$$

We note that a Wald-type test statistic of the null of no interaction $\beta_{GE}=0$ can easily be obtained in a similar fashion, however, the statistic will only have the correct type 1 error level if $\beta_G=0$ also holds.

2 Robustness of profile likelihood

A limitation of the above approach is that when the working regression model (2) is correctly specified, and the independence assumption holds, the test statistic given in equation (7) is not most powerful to detect an alternative hypothesis of non-null genetic associations $\beta_G \neq 0$ or $\beta_{GE} \neq 0$. In order to fully exploit the hypothesis of gene-environment independence, the profile likelihood approach of Chatterjee and Carroll (2005) may be adopted. Under the additional assumption that the disease is rare in the population, the approach entails maximizing the so-called profile likelihood function

$$\prod_i \frac{\text{odds}_{D|G,E,L}(G_i, E_i; \theta)^{D_i} \Pr\{G_i; p\}}{\sum_{d,g} \text{odds}_{D|G,E,L}(g, E_i; \theta)^d \Pr\{G = g; p\}} \quad (8)$$

with respect to θ and p , where $\text{odds}_{D|G,E,L}(G_i, E_i; \theta)$ is defined in equation (2) and $\Pr\{G\} = p^G (1-p)^{1-G}$ is the genotype law in the underlying population.

In the appendix, we show that the corresponding profile maximum likelihood estimator $\bar{\theta}$ shares the partial robustness property of ordinary logistic regression, in the sense that the profile maximum likelihood estimator of (θ_2, θ_3) is guaranteed to converge to $(\beta_G, \beta_{GE}) = (0, 0)$ under the null hypothesis of no genetic effects, even if the model for the odds ratio association $\text{or}(E|G=0)$ relating the continuous exposure variable to the disease outcome among non-carriers is incorrectly specified, provided that G and E are independent in the underlying population. Thus, as in the case of ordinary logistic regression, a Wald-type test statistic of the null hypothesis of no genetic effects is recovered by combining the profile maximum likelihood estimator of (θ_2, θ_3) with Huber's sandwich estimator of its covariance matrix as in equation (6).

3 Final Remarks

One should note that, in the absence of model mis-specification, the profile likelihood test statistic will in general be far more powerful in detecting an alternative hypothesis of a non-null genetic association compared to ordinary logistic regression analysis, provided that the hypothesis of gene-environment independence is satisfied. Otherwise, if the genetic variant and the environmental factor are related in the underlying population, the robustness under the null no longer applies, and both methodologies are vulnerable to bias both under null and alternative hypotheses related to genetic effects (Lindstrom et al, 2009). Furthermore, under

gene-environment dependence, the profile likelihood approach will remain biased even if model (2) is correct (Chatterjee and Carroll, 2005).

It is easily verified that all of the previous results also hold if categorical confounding factors L are accounted for in the model, provided that the corresponding adjusted regression models are fully saturated in L under the null. However, when as will often happen in practice, L is multivariate or includes continuous components, saturated models (with respect to L) may not be feasible. Unfortunately, in such settings, as illustrated in a simulation study in Tchetgen Tchetgen (2010), mis-specification of the effect of exposure generally leads to biased maximum likelihood estimates of (β_G, β_{GE}) when either ordinary logistic regression or the profile likelihood method is used, even if L is correctly entered in the regression models. This bias can in turn result in incorrect inferences of genetic effects. As an alternative approach, Tchetgen Tchetgen (2010) developed a novel constrained retrospective regression framework which does not require the specification of a parametric model for the effect of exposure on the risk of disease, and recovers efficient estimates of genetic effects in the presence of gene-environment interaction and conditional independence given L . Therefore, whereas standard logistic regression and the profile likelihood approach may make incorrect inferences when the effect of exposure is mis-specified, an analysis based on the retrospective regression approach of Tchetgen Tchetgen (2010) is completely immune to this form of mis-specification whether or not the null hypothesis of no genetic effects holds.

References

1. Chatterjee N, Carroll RJ. Semiparametric Maximum-Likelihood Estimation Exploiting Gene-Environment Independence in Case-Control Studies. *Biometrika*. 2005; 92:399–418.
2. Huber, PJ. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press; 1967. The behavior of maximum likelihood estimates under nonstandard conditions; p. 221-223.
3. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect disease susceptibility loci. *Human Heredity*. 2007 Feb 2; 63(2):111–119. [PubMed: 17283440]
4. Lindström S, Yen Y-C, Spiegelman D, Kraft P. The Impact of Gene-Environment Dependence and Misclassification in Genetic Association Studies Incorporating Gene-Environment Interactions. *Hum Hered*. 2009; 68:171–181. [PubMed: 19521099]
5. Rosenblum M, van der Laan M. Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models. *Biometrics*. 2009; 65(3):937–945. [PubMed: 19210739]
6. Tchetgen Tchetgen E. Robust Detection of Genetic Associations in the Presence of Gene-Environment Interaction and Independence. 2010 Under review.

APPENDIX

Let

$$\Pr^{\dagger}(D, G|E) = \frac{\text{odds}_{D|G, E, L}(G, E; \theta)^D \Pr(G; \bar{p})}{\sum_{d, g} \text{odds}_{D|G, E, L}(g, E; \theta)^d \Pr(G = g; \bar{p})}$$

The profile maximum likelihood estimator $(\bar{\theta}, \bar{p})$ solves the score equation of (8):

$$\sum_i \left\{ [1, G_i, G_i E_i, E_i]^T D_i - \sum_{g,d} [1, g, g E_i, E_i]^T d \Pr^\dagger(D = d, G = g | E; \bar{\theta}, \bar{p}) \right\} = 0 \quad \sum_i (G_i - \Pr^\dagger(G = 1 | E; \bar{\theta}, \bar{p})) = 0$$

Let $\theta^\dagger$ denote the limit point (in probability) of $\bar{\theta}$. Therefore, it suffices to verify that $\theta_2^\dagger = 0, \theta_3^\dagger = 0$ solve the population score equation

$$E \left\{ [1, G, GE, E]^T D - \sum_g [1, g, gE, E]^T \Pr^\dagger(D = 1, G = g | E; \theta^\dagger, p^\dagger) \right\} = 0 \quad E(G - \Pr^\dagger(G = 1 | E; \theta^\dagger, p^\dagger)) = 0$$

However, when $\theta_2^\dagger = 0, \theta_3^\dagger = 0$, we obtain $\Pr^\dagger(G = 1 | E; \theta^\dagger, p^\dagger) = p^\dagger = p$ under the null hypothesis. Next, note that

$$\begin{aligned} & \left\{ [1, G, GE, E]^T D - \sum_g [1, g, gE, E]^T d \Pr^\dagger(D = d, G = g | E; \theta^\dagger, p) \right\} = \{ [1, G, GE, E]^T (D - \Pr^\dagger(D = 1 | G, E; \theta^\dagger)) \} \\ & + [1, G, GE, E]^T \Pr^\dagger(D = 1 | G, E; \theta^\dagger) - \sum_g [1, g, gE, E]^T \Pr^\dagger(D = 1 | G = g, E; \theta^\dagger, p) \Pr^\dagger(G = g | E; \theta^\dagger, p) \\ & = \{ [1, G, GE, E]^T (D - \Pr^\dagger(D = 1 | G, E; \theta^\dagger)) \} + \{ [1, 1, E, E]^T \Pr^\dagger(D = 1 | G = 1, E; \theta^\dagger) - [1, 0, 0, E]^T \Pr^\dagger(D = 1 | G \\ & = 0, E; \theta^\dagger) \} (G - p) \end{aligned}$$

Therefore, under the independence assumption

$$E \left\{ [1, G, GE, E]^T D - \sum_g [1, g, gE, E]^T \Pr^\dagger(D = 1, G = g; \theta^\dagger, p) \right\} = E \{ [1, G, GE, E]^T (D - \Pr^\dagger(D = 1 | G, E; \theta^\dagger)) \}$$

which can be handled as in the case of ordinary logistic regression, thus proving that $(\theta_2^\dagger, \theta_3^\dagger) = (\beta_G, \beta_{GE}) = (0, 0)$ solve the population score equation.