Commentary

On the Interpretation, Robustness, and Power of Varieties of Case-Only Tests of Gene-Environment Interaction

Eric J. Tchetgen Tchetgen*

* Correspondence to Dr. Eric J. Tchetgen Tchetgen, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115 (e-mail: etchetge@hsph.harvard.edu).

Initially submitted March 29, 2010; accepted for publication September 27, 2010.

The author addresses issues related to interpretation of the parameters in the proposed case-only linear regression approach to testing for gene-environment interaction recently considered by Clarke and Morris (Am J Epidemiol. 2010;171(4):498–505). He also considers the robustness of their likelihood ratio test to violation of the assumption of Gaussian regression residuals under the null hypothesis of no interaction; shows how their approach can be extended to a more general class of regression models; and derives the optimal interaction test statistic for this class of regression models. Finally, the author briefly discusses case-only doubly robust methods recently proposed by Tchetgen Tchetgen and Robins (Biometrics. 2010;Mar 11).

epidemiologic methods; genotype-environment interaction; likelihood ratio; regression analysis

Abbreviations: CI, confidence interval; LRT, likelihood ratio test; OR, odds ratio.

In a recent article, Clarke and Morris (1) considered 6 types of designs for tests of association in gene-environment (G-E) interaction. Specifically, they modeled, using cases only, the relation between a continuous environmental factor $E$ and a multiallelic genetic factor $G$, by modeling $E$ as the response variable in the first 3 designs and, subsequently, as the exposure variable in the next 3 designs. They proposed likelihood ratio tests (LRTs) of interaction under the assumption of normal residuals when $E$ entails the outcome, and they implemented generalized logistic LRTs to handle cases where $G$ is the outcome. They noted that “more work will be required to interpret the estimated coefficients from these case-only regressions in terms of risk ratios and other measures of effect size” (1, p. 504).

In this commentary, I provide a formal framework for interpreting the parameters of the proposed linear regression approach. In addition, I consider the robustness of the LRT to violation of the normal assumption under the null hypothesis of no interaction; furthermore, I show how the approach of Clarke and Morris (1) can be extended to a more general class of regression models, and I derive the optimal interaction test statistic for this class of regression models. Next, I briefly illustrate the new method with data from an Israeli study of the interaction between parity and breast cancer genes $BRCA1$ and $BRCA2$ in their effects on the risk of ovarian cancer (2). Finally, I briefly discuss the relation between the work of Clarke and Morris (1) and the case-only methods recently proposed by Tchetgen Tchetgen and Robins (3).

**INTERPRETATION OF REGRESSION PARAMETERS**

Similar to Clarke and Morris (1), suppose $E$ is normally distributed among cases, with mean

$$\beta_0 + \mu(G)^T\beta_1$$

and variance $\sigma^2$, where $\mu(G)$ corresponds to a particular choice of genotype coding (i.e., dominant or additive coding). Note that unlike Clarke and Morris (1), I do not assume that $\sigma^2 = 1$. Under this normal model, one can verify, using the invariance property of the odds ratio (OR) function

$$\text{OR}(g,e) = \frac{\Pr(G = g|E = e)\Pr(G = 0|E = 0)}{\Pr(G = 0|E = e)\Pr(G = g|E = 0)}$$

relating $E$ to $G$ among cases, that
OR\left( {g,e;\beta_1,\sigma^2} \right) = \exp\left\{ \frac{\beta_1 e g}{\sigma^2(1)} + \frac{1}{\sigma^2(1)} - \frac{1}{\sigma^2(0)} \right\} \exp\left\{ \frac{1}{2\sigma^2(1)} - \frac{1}{2\sigma^2(0)} \right\} e^{2g} \right\}.

This formula implies that 1) under heteroscedastic error, a test of interaction which entails only testing the hypothesis \( \beta_1 = 0 \) is guaranteed to have power no greater than the type 1 error level to detect any alternative of a nonnull interaction of the form \( \beta_1 = 0 \) and \( \sigma^2(1) \neq \sigma^2(0) \) and 2) a more powerful case-only test of interaction can be obtained either by fitting a logistic regression of \( G \) on \( E \) which includes both a linear and a quadratic term in \( E \) or by fitting the heteroscedastic normal regression model. In the absence of model misspecification, the second option should generally be more powerful, as it makes the additional normality assumption.

Interestingly, one can extend results similar to those of equation 2 beyond simple linear regression with normal homoscedastic error. In fact, a related result was recently established when \( E \) is a Poisson count with mean parameter \( \exp(\gamma_0 + \mu(G)T\gamma_1) \) (4). Similar to the normal case, Osio (4) showed that

\[ \text{OR}(g,e;\gamma_1) = \exp\{\mu(g)^T e\gamma_1\}, \]

so parameters of the case-only log-linear regression of \( E \) on \( G \) can be interpreted as the parameters on the logarithmic scale of multiplicative gene-environment interaction. This result further generalizes to a setting in which \( E \) has a conditional density given \( G \), which I denote by \( f(E|G;\psi_0) \), of the exponential form

\[ h(E)\exp\left[ E\mu(G)^T \psi_0 - c(\mu(G)^T \psi_0) \right], \]

where \( h(E) \) may depend on additional unknown parameters, \( \text{OR}(E,G) = \text{OR}(E,G;\psi_0) = \exp\{E\mu(G)^T \psi_0\} \), and \( c(\mu(G)^T \psi_0) \) is the normalizing constant. In this parameterization, \( \psi_0 \) is the log odds ratio parameter relating \( E \) to \( G \); thus, it constitutes the gene-environment interaction parameter of interest. Statistical inference on parameters of this model is readily obtained by means of the standard maximum likelihood method; further detail is provided below, as well as in the Appendix.

**ROBUSTNESS UNDER THE NULL HYPOTHESIS**

Clarke and Morris (1) investigated the type 1 error rate of their test statistics in a variety of settings via a simulation study. They concluded that their case-only test statistics have a correct type 1 error rate under the null hypothesis of no interaction and that this remains true for the linear regression approach even when the normality assumption does not hold. In the Appendix, I formally prove their empirical observation to be theoretically sound in the linear regression framework. Specifically, when the interaction is null and therefore \( G \) and \( E \) are independent among cases, the least-squares estimator of \( \beta_1 \) is guaranteed to converge (in probability) to zero. Furthermore, even when the normality assumption is incorrect, the LRT statistic will, in large samples, have the correct chi-squared distribution.

In light of the above results, under the assumption that \( E \) follows a density of an exponential form \( f(E|G;\psi_0) \) for a specific choice of \( \psi_0 \), one may test the null hypothesis of no interaction by implementing the conditional (on \( G \)) LRT statistic

\[ 2 \left\{ \log f(E|G;\hat{\psi}) - \log f(E|G;0) \right\} - \chi_k^2 \]

under the null hypothesis,

where \( k \) is the dimension of \( \psi_0 \), \( \hat{f}(E|G;\hat{\psi}) = f(E|G;\hat{\psi},\hat{h}) \) is the conditional (on \( G \)) maximum likelihood estimator of \( f(E|G;\psi_0) \), and \( \hat{f}(E|G;0) = f(E|G;0,\hat{h}) = h(E) \) is the conditional (on \( G \)) maximum likelihood estimator of \( f(E|G;\psi_0) \), with \( \psi_0 \) restricted to equal zero. One can show, using an argument similar to the normal case, that the maximum likelihood estimator \( \hat{\psi} \) will generally converge to zero under the null. However, when \( f(E|G;\psi_0) \) is no longer Gaussian, the LRT may no longer be robust to model misspecification, and its distribution may substantially deviate from the expected chi-squared distribution, often resulting in inflated or deflated type 1 error rates. For this reason, instead of the LRT, we propose use of the following Wald-type test statistic:

\[ n \hat{\psi}^T \Sigma^{-1} \hat{\psi} - \chi_k^2 \]

under the null hypothesis of no interaction, which has the correct type 1 error rate under the null, provided that \( \Sigma \) converges (in probability) in large samples to
the covariance of the scaled (by \( n^{1/2} \)) estimator \( \hat{\psi} \) irrespective of model misspecification. For instance, Huber’s approach (5) may be adopted here to construct a sandwich variance estimator \( \hat{\Sigma} \) which satisfies this latter property. For this choice of \( \hat{\Sigma} \), in addition to being robust under the null, the proposed test statistic is, in large samples, equivalent to the LRT in the absence of model misspecification and is therefore optimal under the model. This is because when the model is correct, \( \hat{\Sigma}^{-1} \) converges to the information matrix for \( \psi_0 \), and \( \hat{\psi} \) is the maximum likelihood estimator.

DATA EXAMPLE: STUDY OF G-E INTERACTION FOR OVARIAN CANCER

Below, I briefly illustrate this new method in an analysis of data from a population-based case-control study based on all ovarian cancer patients identified in Israel between March 1, 1994, and June 30, 1999 (2). In that study, Modan et al. (2) selected 2 controls per case from the central population registry, matching on age within 2 years, area of birth, and place and length of residence. Blood samples were collected in both cases and controls and were used to test for the presence of mutations in the 2 major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. Additional data were collected on reproductive and gynecologic history, such as parity, number of years of oral contraceptive use, and gynecologic surgery. The main objective of the study was to examine the interplay between the BRCA1/2 genes and known reproductive/gynecologic risk factors for ovarian cancer. To test for interactions between reproductive risk factors and BRCA1/2 in their effects on the risk of ovarian cancer, Modan et al. (2) performed the unadjusted case-only analysis of interaction of Piegorsch et al. (6), under an assumption that genetic variants and the environmental factor were unconditionally independent in the population.

In this reanalysis, I illustrate the new case-only method developed here by ignoring data on controls. Specifically, using 832 cases who did not undergo bilateral oophorectomy, were interviewed for risk factor information, and were successfully tested for BRCA1/2 mutations, I show how to estimate the interaction between the dichotomous variable \( G \), representing a person’s BRCA1/2 mutation status (G), and a count variable \( E \), representing a woman’s parity. Under the unconditional independence assumption considered by Modan et al. (2), the standard case-only 1-df logistic LRT of a null odds ratio association between \( G \) and \( E \) yields a \( P \) value of 0.0052, with the corresponding maximum likelihood estimator of \( \hat{\beta}_{GE} = -0.12 \) (95% confidence interval (CI): \(-0.20, -0.03\)). Under the additional assumption that among cases parity follows a Poisson distribution within levels of \( G \), results derived above imply that a valid test of the null hypothesis of no interaction can be obtained via a case-only Poisson LRT of a null log-linear regression of the mean of \( E \) on \( G \), which yields a \( P \) value of 0.0014 and a maximum likelihood estimator of \( \hat{\beta}_{GE} = -0.15 \) (95% CI: \(-0.25, -0.05\)). A \( P \) value of 0.0012 was obtained for the corresponding Wald-type test statistic given above (95% CI: \(-0.24, -0.06\)). Although the results appear to be consistent across the 3 methods, suggesting that the Poisson model provides a reasonable fit to these data, these analyses should be interpreted with caution, since they fail to adjust for additional important factors (3, 7) and are given here solely as an illustration of the methods.

CONCLUSION AND EXTENSIONS

In a recent article, Tchetgen Tchetgen and Robins (3) discussed general case-only likelihood-based methods, including the methods discussed above. Their generalization is of interest, as it recovers the test statistics discussed by Clarke and Morris (1) and further allows the analyst to implement case-only test statistics tailored to enhance power for detecting specific alternative hypotheses of interest. In addition to likelihood-based methods, Tchetgen Tchetgen and Robins also provide a semiparametric approach in the form of a doubly robust case-only method that is relevant to the setting considered by Clarke and Morris and is of particular interest when additional covariates, say \( L \), are also modeled in the case-only regression. Additional covariates will often be included in the case-only regression to control for confounding, as well as to guarantee that gene-environment independence holds within \( L \)-strata of the underlying population (7), a weaker and often more appropriate assumption than assuming gene-environment unconditional independence as considered by Clarke and Morris. The robustness property of the likelihood approach discussed above will not hold in general when such covariates are present in the model, unless one further assumes that the functional form of the regression is correct. In the doubly robust approach, together with a model for the odds ratio, the analyst posits a model for the density of \( E \) given \( G = 0 \) and \( L \) and a model for the density of \( G \) given \( E = 0 \) and \( L \), but remarkably, only 1 of these latter 2 models must be correct in order to obtain an unbiased (more precisely consistent) estimator of the odds ratio parameters (i.e., the interaction parameter of interest) and thus a valid test of gene-environment interaction. We emphasize that under double robustness, the analyst need not know which of the 2 models is correct. Thus, the approach effectively combines the 2 types of models considered by Clarke and Morris, one where \( G \) is the outcome and one where \( E \) is the outcome, into a single approach that offers 2 chances to be correct in one’s inferences, instead of the single chance offered by each of the methods if they are implemented separately.

ACKNOWLEDGMENTS

Author affiliation: Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts.

The author thanks Dr. Tyler VanderWeele for bringing the work of Clarke and Morris (1) to his attention.

Conflict of interest: none declared.
REFERENCES


APPENDIX

Additional technical detail on the exponential family model

According to standard theory for the exponential family model (8), the mean of $E$ given $G$ is given by

$$E(E|G=g) = c_1 \mu(G)^T \psi_0,$$

where $c_1(x)$ is the derivative of $c(x)$ with respect to $x$. In the normal example, $h(E) = N(\beta_0, \sigma^2)$ and

$$c_1(\mu^T(G) \psi_0) = \beta_0 \mu(G)^T \psi_0 + \{\mu(G)^T \psi_0\}^2 \sigma^2/2,$$

so that

$$E(E|G=g) = c_1(\mu(G)^T \psi_0) = \beta_0 + \mu(G)^T \psi_0 \sigma^2$$

as previously established, whereas in the Poisson example, $h(E) = 1/E!$,

$$c_1(\mu(G)^T \psi_0) = \exp(\mu(G)^T \psi_0)$$

and

$$E(E|G=g) = c_1(\mu(G)^T \psi_0) = \exp(\mu(G)^T \psi_0).$$

Simple expressions for $h$, $c$, and $\hat{c}$ are similarly available for a number of other widely used parametric models, including the gamma, binomial, and negative-binomial distributions. Alternatively, the analyst may specify a functional form for $h$ to use in data analysis, with the sole restriction that the corresponding $c$ remains finite. However, a closed-form expression for $c$, $f(E|G; \psi_0)$, and $E(E|G=g)$ may no longer be available under a given choice of $h$, and numerical integration may be required.

Proof of validity of the Gaussian likelihood ratio test under a misspecified model

The ordinary least squares $\hat{\beta}_1$ converges in probability to $\beta_1$, which solves the population normal equation

$$E\left\{\left(\frac{1}{\mu(G)}\right)\{E(E|G) - \beta_0 + \mu(G)^T \beta_1\}\right\} = 0,$$

where $\beta_0 = 0$ and $\beta_1 = E(E)$ solves the equation, since, by the null hypothesis, $E(E|G)$ is constant. According to equation 3.4 of Kent’s (9) theorem 3.1, it suffices to show that under the null hypothesis of no gene-environment interaction, the asymptotic covariance matrix of the scaled (by $n^{1/2}$) maximum likelihood estimator of $\beta_1$ is the same regardless of whether or not the normal model is correct. The result holds because the covariance is equal to $\sigma^2 \{\text{Cov}(\mu(G))\}^{-1}$ in both cases.