Rare disease and other assumptions in statistical analysis of genetic data: the good, the bad, and the ugly

Shili Lin*
The Ohio State University, Columbus, OH, USA - shili@stat.osu.edu

Jingyuan Yang
The Ohio State University, Columbus, OH, USA - summerice0510@gmail.com

Abstract

Statistical applications in genetics, pioneered by R. A. Fisher, have been instrumental in the discoveries of genes and genetic variants predisposed in genetic diseases. To facilitate such applications, various assumptions are frequently made, although their effects are sometimes not well investigated, or worse, the assumptions are "forgotten" and therefore not checked before the application of a statistical procedure. Population genotype frequency distribution, such as Hardy-Weinberg equilibrium (HWE), is often assumed in many statistical procedures, so is the rare disease assumption. One particular problem where such assumptions are made is in the investigation of maternal (genotype) effect in the context of genome-wide association study (GWAS). Maternal effect refers to the phenomenon that the mother's genotype is being expressed in her child's phenotype, regardless of whether the mother's gene was actually passed to the child, leading to an important type of parent-of-origin expression patterns. An effective design for detecting maternal effect is that of case-mother/control-mother, as such a design eliminates the need to genotype sometimes hard-to-recruit fathers. A log-linear model has been proposed to analyze the data to assess the existence of maternal effect, but a number of assumptions are made to avoid overparameterization. They include mating symmetry and allelic exchangeability (weaker conditions than HWE), and the disease being rare. In this paper, we investigate the effects of these assumptions on power, type I error, and biases of parameter estimates based on the log-linear model. We show that *good* results, with an increase in power, may indeed be obtainable under certain situations. However, the results can be *bad*, leading to a substantial amount of bias even if the assumptions are met. When some of the assumptions are violated, the results are *ugly*; there can be severely inflated type I errors and huge biases. In contrast, we offer a logistic model as an alternative approach for detecting maternal effects for data from the same design. Our simulation study show that the type I error rates are well controlled without compromising much power, regardless of whether the assumptions hold or not.

Keywords: Bias and type I error, maternal effect, rare disease assumption, Hardy-Weinberg equilibrium.

1 Introduction

Mendelian expression patterns have been observed in most genes. However, the expressions of some genes appear to exhibit parent-of-origin patterns, which have been increasingly studied in the past decades since the reporting of the first human gene subjected to parental imprinting (Giannoukakis *et al.*, 1993). In the literature, the terms parent-of-origin effect and genomic imprinting have frequently been used interchangeably, but in recent years, biologists have reported maternal effect as an alternative cause of parent-of-origin patterns (Wittkopp *et al.*, 2006, Hager *et al.*, 2008). Phenotypes are usually considered as a result of interaction between genetic and environmental factors. The prenatal environment provided by the mother could contribute to the offspring's phenotype because the expression level of certain genes could be altered by the additional mRNAs passed to the offspring from the mother during pregnancy. Hence, although maternal effect is an important environmental factor, it may also be regarded as genetically based.

Maternal effect has gained much attention in statistical genetics research in an effort to detect its association with complex traits. Weinberg *et al.* (1998) and Weinberg (1999) proposed a method to detect maternal effect and imprinting effect simultaneously using case-parent triads, while Sinsheimer *et al.* (2003) developed a likelihood approach to detect maternal effect and non-inherited maternal allele (NIMA) effect using nuclear families with multiple affected children. More recently, Yang and Lin (2013) and Han *et al.* (2013) proposed a robust partial likelihood approach to detecting parent-of-origin effects. Family-based association studies are the natural platforms to study maternal effect, as family data are indispensable for detecting any parent-of-origin effects.

In Shi *et al.* (2008), the authors considered a case-mother/control-mother design to study maternal effect assuming there is no imprinting effect. This design eliminates the need to genotype the sometimes hard-to-recruit fathers needed in case-parent triad designs or in other family-based association studies. Based on this design, maternal effect can be detected using either a logistic or a log-linear model. Although the latter might achieve a slightly higher power by accommodating parameter constraints from three levels of mating frequency assumptions arising from Mendelian inheritance, mating symmetry and parental allelic exchangeability, as we will see below, its use can lead to severe biases in parameter estimates if such assumptions are violated. Specifically, in this paper, we first investigated whether the rare disease assumption is sufficient to warrant the use of population-based child-mother genotype frequency distribution as that of control-mother pairs. Then we evaluated the relative merits of the use of logistic model versus the log-linear model with constraints. Further, the mating symmetry assumption was also examined for tests based on the more traditional triad design.

2 Rare disease assumption

Consider the following log-linear model:

$$P(D=1 \mid M, C) = \delta R_1^{I(C=1)} R_2^{I(C=2)} M_1^{I(M=1)} M_2^{I(M=2)},$$

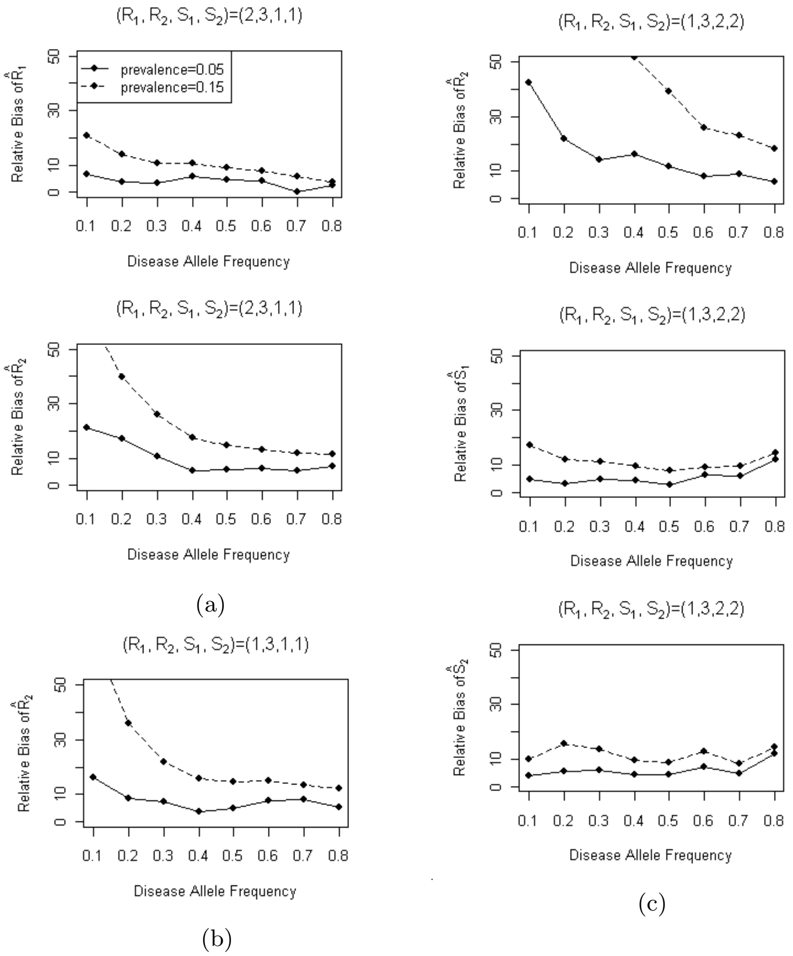
where $D = 1$ indicates that a child is affected with a disease, M and C denote the numbers of disease alleles carried by mother and child, respectively, δ is the phenocopy rate of the disease (i.e., having disease without carrying any disease allele by the child or the mother), R_1 and R_2 are the relative risks due to one or two copies of disease alleles carried by the child, respectively, and I is the usual indicator function taking the value of 0 or 1. Further, S_1 and S_2 are the relative risks due to one or two copies of the disease alleles carried by the mother, respectively, which denote the maternal effects. Although the frequencies in Table 1 of Shi *et al.* (2008) were referred to as expected frequencies of control-mother pairs, they are actually expected frequencies of any child-mother pairs with particular genotype combinations regardless of whether the child is affected or not. Considering the retrospective nature of the case-mother/control-mother design, the fact that the child in a control-mother pair is unaffected should be taken into account by multiplying the unaffected probabilities to the expected frequencies.

The corrected expected frequencies of control-mother pairs are shown in Table 1 of this paper, which, as can be seen, differ from those in Table 1 of Shi *et al.* (2008) by the unaffected probabilities. Unless the

Table 1: Actual expected frequencies of control-mother pairs without approximation[a]

	$C=0$	$C=1$	$C=2$
$M=0$	$B(1-\delta)[\mu_{00}+(1/2)\mu_{01}]$	$B(1-\delta R_1)[(1/2)\mu_{01}+\mu_{02}]$	0
$M=1$	$B(1-\delta S_1)[(1/2)\mu_{10}+(1/4)\mu_{11}]$	$B(1-\delta R_1 S_1)(1/2)[\mu_{10}+\mu_{11}+\mu_{12}]$	$B(1-\delta R_2 S_1)[(1/4)\mu_{11}+(1/2)\mu_{12}]$
$M=2$	0	$B(1-\delta R_1 S_2)[\mu_{20}+(1/2)\mu_{21}]$	$B(1-\delta R_2 S_2)[\mu_{22}+(1/2)\mu_{21}]$

[a] Note that C and M are the number of disease allele(s) carried by the child and mother in the control-mother pairs, respectively; μ_{mf} denotes the population frequency of parental pairs in which the mothers carry m copies and the fathers carry f copies of the disease allele; δ is the phenocopy rate of the disease in the source population; R_1 and R_2 are risks (relative to the phenocopy rate) due to one and two disease allele(s) carried by the offspring; S_1 and S_2 are risks (relative to the phenocopy rate) due to the maternal effect of one and two disease allele(s) carried by the mother; B is a normalizing constant included to ensure that the expected frequencies will sum to the total number of control-mother pairs, which is in fact the (unconditional) probability that a child random chosen from the population is unaffected.



(a)

(b)

(c)

Figure 1: Relative estimation biases of parameters with two disease prevalence, 0.05 and 0.15, under three disease risk settings. (a):(R_1, R_2, S_1, S_2)=(2, 3, 1, 1) and the biases for R_1 and R_2 are plotted; (b): (R_1, R_2, S_1, S_2)=(1, 3, 1, 1) and the estimation bias for R_2 is plotted; (c) (R_1, R_2, S_1, S_2)=(1, 3, 2, 2) and the biases for R_2, S_1, and S_2 are plotted. Note that for a disease with known prevalence and fixed R_1, R_2, S_1, S_2, the phenocopy rate (used in table 1) and the disease allele frequency is monotonically related, and thus the figures could also be plotted in terms of the phenocopy rate.

conditional probability of being unaffected in each (M, C) cell is close to 1, approximating the frequencies of child-mother pairs using the frequencies of control-mother pairs would be inaccurate. The justification for the use of this approximation is that the rare disease assumption is in place (Shi *et al.*, 2008). This assumption implies that the unaffected probability is close to 1 in the whole population, that is, the normalizing constant B in table 1 is approximately 1. However, even if the disease is indeed rare, the correction factor (i.e., conditional probability) in some of the cells may not necessarily be close to 1. In particular, the $(M = 2, C = 2)$ cell often cannot be approximated well.

To illustrate the fact that such inaccurate estimation of control-mother pair frequencies may lead to large biases in the parameter estimation even when the disease is rare, we simulated 150 case-mother pairs and 150 control-mother pairs under the same settings of relative risks (R_1, R_2, S_1 and S_2) as in the literature (Shi *et al.*, 2008) to facilitate comparison, but fixed the prevalence of the disease at 0.05 and 0.15 in our simulation. One thousand replicates of the simulated data were analyzed using the logistic model, which is mathematically equivalent to the log-linear model without any constraint (Agresti, 1990), but differs from the constrained log-linear model. Relative bias of $\hat{R}_1$ is defined as $\frac{\hat{R}_1 - R_1}{R_1} \times 100\%$; similar definitions apply

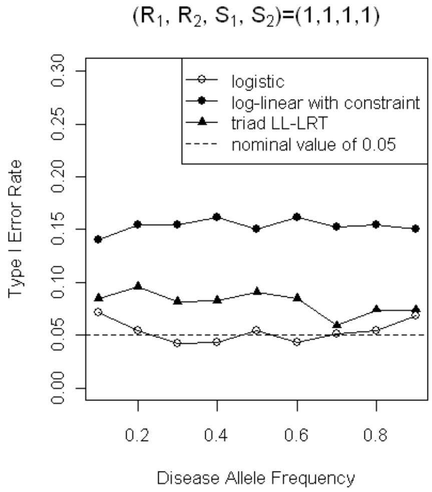


Figure 2: Type I error rates under the null setting for several methods.

to the other parameters. Figure 1 plots the relative estimation biases of the parameters with a true value larger than 1. As we can see from the figure, the biases are smaller for a rare disease (prevalence = 0.05) than for a common disease (prevalence = 0.15). However, appreciable biases still exist even if the disease is rare, especially when the disease allele frequency is small. For example, the two plots in Figure 1a are for the estimation biases of R_1 (=2) and R_2 (=3) (but not for S_1 and S_2 since they both have a null value of 1), which show that the relative estimation biases can be as large as 20% when the disease allele frequency is 0.1 under the rare disease setting. This illustration serves as a cautionary note to bear in mind that estimation bias may still be an issue even if the rare disease assumption holds. An alternative would be to use the corrected frequencies given in Table 1 when using this study design.

3 Logistic Model

The use of log-linear model can incorporate parameter constraints arising from mating frequency assumptions and hence can potentially achieve a higher power. However, its sensitivity to departure from these assumptions has not been ascertained. As an alternative, logistic model may be used instead without making use of these assumptions, although potentially at the expense of power loss. Poisson (log-linear) model and logistic regression model are mathematically equivalent in their basic forms, that is, the equivalence of these two models is achieved when no constraints are imposed on the model parameters. However, when assumptions are imposed on the log-linear model in the analysis, the maximum likelihood estimates of the parameters from these two models will be different, leading to different type I error rates and different power.

To investigate the effects of the three mating frequency assumptions (namely, Mendelian inheritance, mating symmetry, and parental allelic exchangeability) on the log-linear and logistic models, we simulated case-mother/control-mother pairs under the null setting, (R_1, R_2, S_1, S_2)=(1,1,1,1), but without allelic exchangeability and mating symmetry assumptions by introducing the inbreeding factor in the genotype frequencies (Weir, 1996). Specifically, the frequencies of the fathers and the mothers carrying 0, 1, and 2 copies of the disease allele are respectively $(1-p)^2(1-F)+(1-p)F$, $2p(1-p)(1-F)$, and $p^2(1-F)+pF$, where p is the disease allele frequency. The parameter F is the inbreeding factor, which is set to be 0.15 and 0.1 for mothers and fathers, respectively. We fitted a logistic model and a log-linear model with the constraints. Type I error rates of these two models under the whole range of allele frequencies are plotted in Figure 2. Compared to the nominal level of 0.05 (dashed line), type I error rates of the logistic model (lines with empty circles) are well controlled, whereas the type I error rates of the log-linear model (lines with filled circles) with the constraints are greatly inflated.

On the other hand, although imposing the constraints in the log-linear model generally leads to a power

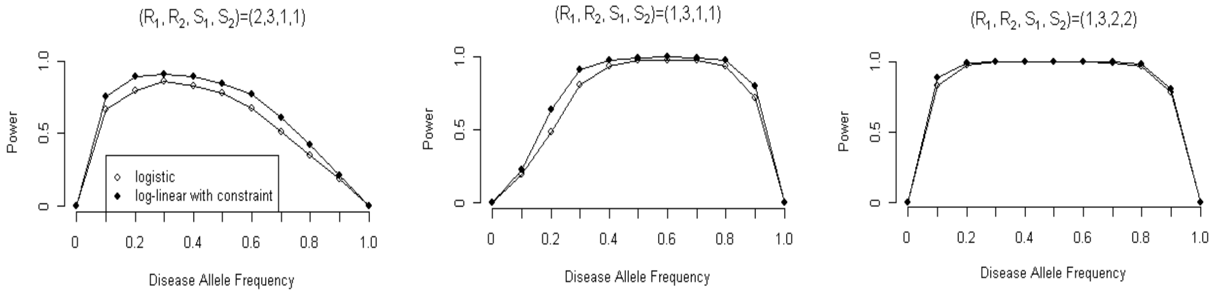


Figure 3: Power of logistic model and log-linear model with constraint.

gain, the gain can be quite small nonetheless. Although Figure 1 of Shi *et al.* (2008) shows a large difference between the power curves of the logistic model and the log-linear model with the allelic exchangeability constraint, the differences in power between the two models as shown are in fact the same as shown in our Figure 3. The differences in their appearance is due to the difference in the scales used in the two drawings. In Figure 1 of Shi *et al.* (2008), the left-hand-side y-axis is the non-centrality parameter of a chi-squared distribution, which does not translate linearly in terms of power. As we can see from the right-hand-side y-axis, power is greatly stretched in the upper end to match the linear scale of non-centrality parameter, which stretches out the power curves even though there is little difference. On the other hand, our Figure 3 plotted the same two power curves (logistic versus log-linear with allelic exchangeability and using the same underlying disease model), but according to the linear scale of power. As such, it is seen clearly that the power gain is not substantial. For example, when maternal effect is present, i.e., under the setting (R_1, R_2, S_1, S_2)=(1, 3, 2, 2), the maximum power gain by imposing the parameter constraints is about 0.025 (bottom panel of Figure 3). These results lead us to conclude that a logistic model may be favorable for its stable type I error rate without much compromise in power.

4 Population Frequency Assumption for Triad Design

Mating symmetry is a necessary assumption for most of the maternal and/or genomic imprinting effects detection methods (Weinberg *et al.*, 1998, Weinberg, 1999, Sinsheimer *et al.*, 2003), whereas the assumption is unnecessary when using a logistic model to analyze the data from a case-mother/control-mother design. We simulated 150 case-parent triads under the null setting (R_1, R_2, S_1, S_2)=(1,1,1,1) with mating symmetry being violated (using the aforementioned inbreeding parameters), and use the log-linear likelihood ratio test (LL-LRT) (Weinberg *et al.*, 1998) to detect the maternal effect. Type I error rates of LL-LRT are showed in Figure 2 (lines with triangles), all of which are about 1.5 to 2 times as large as the nominal level of 0.05. By contrast, in a logistic model, parameters involving mating frequencies are canceled out when comparing the case-mother pairs with the control-mother pairs, regardless of whether there is mating symmetry or not. As such, its type I error rates are well controlled even if mating symmetry does not hold.

5 Conclusions

In this paper, we address several issues related to the case-mother/control-mother and the triad designs for detecting maternal effects. Our simulation results show that treating a randomly selected child-mother pair from the general population as a control-mother pair may lead to bad results with considerable biases in the parameter estimates even if the disease is rare (usually defined to be with prevalence less than 0.1). As such, the exact, instead of the approximate, expected frequencies of the control-mother genotype combinations should be used even if the disease is rare. A number of mating frequency assumptions, in particular parental mating symmetry, have been assumed in statistical methods for detecting parent-of-origin effects. Although Mendelian inheritance is a natural assumption, mating symmetry and allelic exchangeability can easily be violated under an assortative mating scheme. Our results indicate that, although there may be good news of some power gain, the type I error rates can be greatly inflated using the log-linear model if some of the assumptions are violated, leading to extremely ugly outcomes. By contrast, the logistic model is robust to potential violation of mating symmetry and other constraints, but without much power reduction in the case that the assumptions indeed hold. The trade-off of a slight power reduction certainly outweighs the potential of a much inflated type I error rate. These results lead us to recommend the use of logistic model (or equivalently, log-linear model without any constraint) for detecting parent-of-origin effects under certain study designs.

References

Agresti, A. (1990). *Categorical data analysis*. John Wiley & Sons.

Giannoukakis, N., Deal, C., Paquette, J., Goodyer, C. G., and Polychron, C. (1993). arental genomic imprinting of the human igf2 gene. *Nature Genetics*, **4**(1), 98–101.

Hager, R., Cheverud, J. M., and Wolf, J. B. (2008). Maternal effects as the cause of parent-of-origin effects that mimic genomic imprinting. *Genetics*, **178**, 1755–1762.

Han, M., Hu, Y.-Q., and Lin, S. (2013). Joint detection of association, imprinting and maternal effects using all children and their parents. *EUROPEAN JOURNAL OF HUMAN GENETICS*, **21**(12), 1449–1456.

Shi, M., Umbach, D. M., Vermeulen, S. H., and Weinberg, C. R. (2008). Making the most of case-mother/control-mother studies. *American Journal of Epidemiology*, **168**(5), 541–547.

Sinsheimer, J. S., Palmer, C. G., and Woodward, J. A. (2003). Detecting genotype combinations that increase risk for disease: The maternal-fetal genotype incompatibility test. *Genetic Epidemiology*, **24**, 1–13.

Weinberg, C. R. (1999). Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet*, **65**, 229–235.

Weinberg, C. R., Wilcox, A. J., and Lie, R. T. (1998). A log-linear approach to case-parent-triad data: Assessing effects of disease genes that act either directly or through maternal effects and that may be subjected to parental imprinting. *Am J Hum Genet*, **62**, 969–978.

Weir, B. S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2006). Parent-of-origin effects on mrna expression in drosophila melanogaster not caused by genomic imprinting. *Genetics*, **173**, 1817–1821.

Yang, J. and Lin, S. (2013). Robust Partial Likelihood Approach for Detecting Imprinting and Mat ernal Effects Using Case-Control Families. *The Annals of Applied Statistics*, **7**, 249–268.