An indirect test of gene-environment interaction for binary trait

Ziang Zhang[1] and Lei Sun[1, 2]

[1]Department of Statistical Science, University of Toronto, Ontario M5S 3G3, Canada

[2]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Ontario M5T 3M7, Canada

Abstract

In genome-wide association studies (GWAS), it is desirable to include and test the interaction effect (GxE) between single-nucleotide polymorphism (SNP,G) and environmental variable (E), to achieve higher power for the detection of causal SNPs. However, accounting for this interaction effect through direct testing is infeasible in most studies, because the information on environmental variable E is not available or is collected with a non-trivial amount of measurement errors. On the other hand, the indirect testing method allows this interaction effect to be detected and hence accounted for without using the information of E. For quantitative traits (Y) that are approximately normally distributed, it has been shown that indirect testing on GxE interaction can be done by testing the heteroskedasticity of Y between genotypes. Therefore, screening out SNPs with strong signal of heteroskedasticity helps to identify potential causal SNPs that should be studied in a more detailed follow-up analysis, when the traits are quantitative. However, when traits are binary, the existing methodology based on testing for the heteroskedasticity between genotypes cannot be generalized for the same problem. In this paper, we proposed a novel methodology for indirect testing of (GxE) interaction effect for binary traits. Through extensive simulation studies using the 1000 Genome Project (1kGP) dataset, we will show that the proposed approach has a well-controlled type I error rate, and satisfactory power compared to the direct testing method. In the end, we will illustrate the use of the proposed method by applying it to the UK biobank dataset for a GWAS study.

1. Introduction

In traditional genome-wide association studies (GWAS), the detection of genetic association between a single-nucleotide polymorphism (SNP, G) and a quantitative or binary trait (Y) is often done by testing on the main effect of G alone. This framework will have inferior power if the true casual SNPs have trivial main effects but strong interaction effects with some environmental variable (E). A known example of such SNP is rs12753193(LEPR), which has been shown to have a strong interaction effect with BMI on C-reactive protein levels but no detectable main effect (Paré et al., 2010).

The main reason that accounting gene-environment interaction effect (GxE) is generally difficult in most GWAS studies is information from the interacting environmental variable E is often missing or collected with non-trivial measurement error (Soave et al., 2015). For quantitative traits (Y) that are approximately normally distributed, it has been shown that ignoring the environmental variable E when it is interacting with a SNP will produce an artificial heteroskedasticity of Y across genotypes (Paré et al., 2010). For this reason, indirect testing methods that detect the GxE interaction by testing the heteroskedasticity of Y have been proposed for quantitative traits (Soave et al., 2015; Soave and Sun, 2017). The indirect testing method enables the detection of significant GxE interaction even if there is no information collected on variable E, and hence allows its effect to be accounted for easily for studies of genetic association between a SNP with a quantitative trait.

However, the approach of testing potential GxE interaction by checking variance of traits across groups will not generalize to the case where traits of interest are binary such as disease affection status, because the variance of binary variable is directly specified by its mean. This places a constraint on how to account for the GxE interaction effect in a GWAS study for a binary trait.

In this paper, we propose a novel methodology for indirect testing of GxE interaction for binary traits by exploring the latent variable framework of the probit regression model.

Using the latent variable framework, we will first show that for a SNP with additive main effect, ignoring the GxE interaction will result in a non-additive genotypic model, and hence propose a method to detect the GxE interaction without using the information of E. Then we will show that under the existence of a suitable auxiliary variable Z, the power of our proposed method can be improved, and the proposed method can be generalized to the case where the main effect of SNP is genotypic instead of additive.

The remainder of this paper is organized as follows. In section 2, we will give more details on the latent variable formulation of logistic/probit regression, and discuss the details of why methods based on the Levene test will not work in this setting. In section 3, we will describe our novel methodology for indirect testing of GxE interaction, and explain how it can be improved if there exists suitable auxiliary variable Z. In section 4, we will show that our proposed approach has a well-controlled type I error rate and satisfactory power even compared to the direct testing method with information of E available, through an extensive set of simulation studies using the 1000 Genome Project (1kGP) dataset (A et al., 2015). In section 5, we will implement our proposed method on the UK Biobank (UKB) dataset, to identify promising SNPs that are likely to be interacting with some unknown environmental variables (Sudlow et al., 2015). We conclude with a discussion in section 6.

2. Preliminaries

2.1. Latent formulation of Generalized linear Model

Let Y be the binary trait of interest, G be the count of minor allele for the SNP of interest (so G can take 0, 1, or 2) and E be the environmental variable. The conventional method to model such binary response will be through the following generalized linear model (GLM):

$$P(Y = 1) = g^{-1}(\beta_0 + \beta_G G + \beta_E E + \beta_{GE} GE) \tag{1}$$

where the function g^{-1} refers to the inverse of link function of the GLM. If a logistic regression is used, g^{-1} will be the logistic function. If a probit regression being used, g^{-1} will be the cumulative distribution function (CDF) of the standard normal distribution.

An equivalent formulation of the model above will be through the latent variable formulation. Define Y^* be a latent variable that cannot be directly observed, generated by the following model:

$$Y^* = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} GE + \epsilon \tag{2}$$

the random error ϵ is assumed to be independent of all the covariates in the model, and can either have logistic distribution or have normal distribution, depending on whether the model 1 is logistic or probit.

The latent variable Y^* is not observable, but it generates the binary observations Y in the following way:

$$Y = \mathbb{I}\{Y^* > 0\} \tag{3}$$

In other words, the response variable Y can be viewed as an indicator variable defined based on the magnitude of the latent variable.

Assuming that the probit model 2 with random error $\epsilon \sim N(0, \sigma_\epsilon^2)$ generates the binary response variable, then the conditional probability of observing a case (i.e. Y = 1) can be computed as:

$$\begin{aligned} \mathrm{P}(Y = 1|G, E) &= \Phi\left(\frac{\mathbb{E}(Y^*|G, E)}{\sqrt{\mathrm{Var}(Y^*|G, E)}}\right) \\ &= \Phi\left(\frac{\beta_0 + \beta_G G + \beta_E E + \beta_{GE} GE}{\sigma_\epsilon}\right) \\ &= \Phi(\tilde{\beta}_0 + \tilde{\beta}_G G + \tilde{\beta}_E E + \tilde{\beta}_{GE} GE) \end{aligned} \tag{4}$$

Where the parameter $\tilde{\beta}$ denotes $\frac{\beta}{\sigma_\epsilon}$. The parameter σ_ϵ will not be identifiable in the model, since any scalar multiplication simultaneously on β and σ_ϵ will yield the same value of $\tilde{\beta}$. Therefore, in probit regression, the regression parameters actually refer to $\tilde{\beta}$ instead of β. In the rest of this work, unless stated otherwise, we will assume the random error has been

properly standardized so $\tilde{\beta} = \beta$ for all the regression parameters.

A significant advantage of the probit model over the logistic model is that if $E \sim N(0, \sigma_E^2)$ is independent of both the random error ϵ and the SNP of interest G, the resulting model conditional on G will still be a valid probit model. This result holds because $E + \epsilon$ will still be a normally distributed random error independent of G. However, if ϵ follows a logistic distribution instead, the same result will not necessarily hold even if E also follows a logistic distribution. Therefore, the resulting model with a missing GxE interaction is not easily tractable if a logistic regression is used instead.

2.2. Indirect testing of interaction effect using Levene's method

The goal of indirect testing of interaction effect is to assess whether $\beta_{GE} = 0$ in model 2, without using information of the environmental variable E. Assume for now that the latent variable Y^* can be observed, and the unknown environmental variable E is generated from $N(0, \sigma_E^2)$, independent of both the SNP G and the random error ϵ. Furthermore, assume that the true underlying model that generates the binary response variable is a probit model, and the random error ϵ follows distribution $N(0, \sigma_\epsilon^2)$.

With this specification, the problem reduces to the indirect testing of interaction effect when the "response" variable is quantative (i.e. Y^* is quantative), instead of binary. As Soave and Sun (2017) have proposed, the indirect testing problem in this case can be done using the notion of generalized Levene test. If the interaction effect $\beta_{GE} \neq 0$, then the previous model 2 can be reduced to the following heteroskedastic linear regression model:

$$Y^* = \beta_0 + \beta_G G + \epsilon_G \tag{5}$$

where the new random error ϵ_G will have its variance dependent on G, in the following way:

$$\mathrm{Var}(\epsilon_G | G = g) = \mathrm{Var}(Y^* | G = g) = (\beta_E \sigma_E + \beta_{GE} g)^2 + \sigma_\epsilon^2 \tag{6}$$

Equation 6 implies that the variance of Y^* will be constant across different genotypic groups if and only if $\beta_{GE} = 0$. Therefore, the Levene-type test proposed in (Soave et al., 2015; Soave and Sun, 2017) can indirectly test for interaction by testing on the hypothesis of constant variance across genotypic groups.

However, the Levene-type method described above will not work for this problem, because the latent variable Y^* is not *observable* in practice. The observable traits Y is linked to $\text{Var}(Y^*|G)$ only through $\mathbb{E}(Y^*|G)/\sqrt{\text{Var}(Y^*|G)}$, and hence the quantity $\text{Var}(Y^*|G)$ will not be identifiable from data of Y.

3. Methods

3.1. Testing based on additivity of SNP effect

Consider the true underlying probit model is model 2, with $\beta_{GE} \neq 0$, where $E \sim N(\mu_E, \sigma_E^2)$ follows the classical *G-E* independence assumption (Lindstrom et al., 2009), and $\epsilon \sim N(0, \sigma_\epsilon^2)$ is independent of both the SNP of interest G and the environmental variable E. Under these assumptions, the conditional mean and variance of Y^* can be computed as:

$$\begin{aligned} \mathbb{E}(Y^*|G) &= \beta_0 + \beta_E\mu_E + (\beta_G + \beta_{GE}\mu_E)G \\ \text{Var}(Y^*|G) &= (\beta_E + \beta_{GE}G)^2\sigma_E^2 + \sigma_\epsilon^2 \end{aligned} \tag{7}$$

Therefore, if the environmental variable E is omitted from the model, the resulting probit model will reduce to the following:

$$\begin{aligned} \text{P}(Y = 1|G) &= \Phi\left(\frac{\mathbb{E}(Y^*|G)}{\sqrt{\text{Var}(Y^*|G)}}\right) \\ &= \Phi\left(\frac{\beta_0 + \beta_E\mu_E + (\beta_G + \beta_{GE}\mu_E)G}{\sqrt{(\beta_E + \beta_{GE}G)^2\sigma_E^2 + \sigma_\epsilon^2}}\right) \\ &= \Phi(\gamma_0 + \gamma_1\mathbb{I}(G = 1) + \gamma_2\mathbb{I}(G = 2)) \end{aligned} \tag{8}$$

Where the parameters γ_0, γ_1 and γ_2 are defined as

$$\begin{aligned}
\gamma_0 &= \frac{\mathbb{E}(Y^*|G=0)}{\mathrm{Var}(Y^*|G=0)} = \frac{\beta_0 + \beta_E\mu_E}{\sqrt{(\beta_E^2\sigma_E^2 + \sigma_\epsilon^2)}} \\
\gamma_1 &= \frac{\mathbb{E}(Y^*|G=1)}{\mathrm{Var}(Y^*|G=1)} = \frac{\beta_0 + \beta_E\mu_E + (\beta_G + \beta_{GE}\mu_E)}{\sqrt{((\beta_E + \beta_{GE})^2\sigma_E^2 + \sigma_\epsilon^2)}} - \gamma_0 \\
\gamma_2 &= \frac{\mathbb{E}(Y^*|G=2)}{\mathrm{Var}(Y^*|G=2)} = \frac{\beta_0 + \beta_E\mu_E + 2(\beta_G + \beta_{GE}\mu_E)}{\sqrt{((\beta_E + 2\beta_{GE})^2\sigma_E^2 + \sigma_\epsilon^2)}} - \gamma_0
\end{aligned} \tag{9}$$

Notice that this model is still a valid probit model, but the effect of SNP G changes from additive to genotypic. If $\gamma_1 = 0.5\gamma_2$, then the simplified probit model is still additive in G, but that will only happen when $\beta_{GE} = 0$.

Therefore, when the effect of G on Y^* is additive, testing the hypothesis $\gamma_1 = 0.5\gamma_2$ is equivalent to testing $\beta_{GE} = 0$. Although in model 2, both G and its interaction with E are assumed to affect Y^* additively, this method will still work if the interaction effect between G and E is non-additive.

3.2. Testing based on the auxiliary variable

In this section, we assume that there exists an auxiliary variable Z in the true generating model, so the latent model 2 can be written as:

$$Y^* = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} GE + \beta_Z Z + \epsilon \tag{10}$$

The auxiliary variable Z has to satisfy the following three properties:

1. Observations of Z are availiable in the dataset.

2. Z has no interaction with the SNP of interest G

3. Z is independent of the random error ϵ

Given such an auxiliary variable Z exists in the dataset, then the conditional probability $\mathrm{P}(Y=1|G,Z)$ can be written as the following:

$$\begin{aligned}\mathrm{P}(Y=1|G,Z) &= \Phi\left(\frac{\mathbb{E}(Y^*|G)}{\sqrt{\mathrm{Var}(Y^*|G)}}\right) \\ &= \Phi\left(\frac{\beta_0+\beta_E\mu_E+(\beta_G+\beta_{GE}\mu_E)G+\beta_Z Z}{\sqrt{(\beta_E+\beta_{GE}G)^2\sigma_E^2+\sigma_\epsilon^2}}\right) \\ &= \Phi(\gamma_0+\gamma_1\mathbb{I}(G=1)+\gamma_2\mathbb{I}(G=2)+\gamma_Z Z+\gamma_{Z1G}\mathbb{I}(G=1)Z+\gamma_{Z2G}\mathbb{I}(G=2)Z)\end{aligned} \tag{11}$$

The parameters $\gamma_0,\gamma_1,\gamma_2$ are the same as in equations 9. The new parameters $\gamma_Z,\gamma_{Z1G},\gamma_{Z2G}$ are defined as:

$$\begin{aligned}\gamma_Z &= \frac{\beta_Z}{\sqrt{\beta_E^2\sigma_E^2+\sigma_\epsilon^2}} \\ \gamma_{Z1G} &= \frac{\beta_Z}{\sqrt{(\beta_E+\beta_{GE})^2\sigma_E^2+\sigma_\epsilon^2}}-\frac{\beta_Z}{\sqrt{\beta_E^2\sigma_E^2+\sigma_\epsilon^2}} \\ \gamma_{Z1G} &= \frac{\beta_Z}{\sqrt{(\beta_E+2\beta_{GE})^2\sigma_E^2+\sigma_\epsilon^2}}-\frac{\beta_Z}{\sqrt{\beta_E^2\sigma_E^2+\sigma_\epsilon^2}}\end{aligned} \tag{12}$$

Assuming that $\beta_Z \neq 0$, the equation above shows that a missing interaction (GxE) creates an *artificial* non-additive interaction between the auxiliary variable Z and the genotypes of the SNP of interest G.

Therefore, if an auxiliary variable Z exists and is known to satisfy the three requirements in 3.2 with non-zero β_Z, one can also test the hypothesis $\gamma_{Z1G}=\gamma_{Z2G}=0$ in order to test $\beta_{GE}=0$. Although it is assumed in model 10 that G has additive effect β_G, it is clear that this methodology will still hold if G has non-additive effect or if GxE interaction is non-additive.

If it is already known that the effect of G should be coded additively in the model, then the above methodology can be incorporated into the methodology of testing additivity proposed in section 3.1, by jointly testing the null hypothesis

$$H_0 : \gamma_1 = 0.5\gamma_2, \gamma_{Z1G}=\gamma_{Z2G}=0$$

This will boost the power of detecting non-zero interaction β_{GE}, by both reducing the variance of random error ϵ and checking whether there are pseudo interactions between Z and G created by the missing GxE interaction.

In the rest of this paper, all the hypotheses described above will be tested through the Wald test. That means, method of testing additivity proposed in section 3.1 will be using one degree of freedom Chi-Square test, the method based on auxiliary variable proposed earlier in this section will be using two degrees of freedom Chi-Square test, and their combination will be using three degrees of freedom Chi-Square test.

4. Simulations

here goes the 1kGP simulation study

5. Examples

Here goes an example of implementing the proposed method for UKbiobank dataset

6. Discussion

Discuss about the current limitations of the proposed method

References

A, A., Abecasis, G., DM, A., RM, D., DR, B., A, C., AG, C., P, D., EE, E., P, F., SB, G., Gibbs, R., ED, G., ME, H., Knoppers, B., JO, K., ES, L., Lee, C., Lehrach, H., and JA, S. (2015). A global reference for human genetic variation. *Nature*, 526:68.

Lindstrom, S., Yen, Y.-C., Spiegelman, D., and Kraft, P. (2009). The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. *Human heredity*, 68:171–81.

Paré, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the women's genome health study. *PLOS Genetics*, 6(6):1–10.

Soave, D., Corvol, H., Panjwani, N., Gong, J., Li, W., Boelle, P.-Y., Durie, P., Paterson, A., Rommens, J., Strug, L., and Sun, L. (2015). A joint location-scale test improves power to detect associated snps, gene sets, and pathways. *American journal of human genetics*, 97:125–138.

Soave, D. and Sun, L. (2017). A generalized levene's scale test for variance heterogeneity in the presence of sample correlation and group uncertainty. *Biometrics*, 73(3):960–971.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):1–10.