Progress report: Detecting interaction with unknown environmental covariate

Ziang Zhang

15/10/2020

1 The Underlying Model:

For binary response variable, it is often assumed that the response variable y_i conditioning on the regressors G_i, Z_i come from a latent model such that:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_G G_i + \beta_Z Z_i + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \end{aligned} \tag{1}$$

The unobserved latent variable Y_i^* determines whether the observed response variable Y_i is 0 or 1. The error term ϵ_i in Y_i^* needs to have a completely known distribution, which can be $\mathrm{N}(0,1)$ for the model to become a probit model, or a logistic distribution with mean 0 and variance 3.28 for the model to become a logistic regression model.

Here the regressor G_i represents the allele of interest, and the regressor Z_i is any regressor that can be non-genetic. For now on, we will assume the model is probit for simplicity, unless otherwise indicated.

Similarly, we can have a Genotypic Model defined as:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_{G1} I(G_i = 1) + \beta_{G2} I(G_i = 2) + \beta_Z Z_i + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \end{aligned} \tag{2}$$

The Genotypic Model has higher degree of freedom than the additive model due to the extra regression parameter.

1.1 When the true model does contain gene-environment interaction

Assume for simplicity that E_i the environmental variable has a normal distribution with mean μ_E and variance σ_E^2, and suppose that the true underlying model is:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_G G_i + \beta_Z Z_i + \beta_E + E_i + \beta_{G\times E} G_i \times E_i + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \\ \epsilon_i &\sim \mathrm{N}(0,1) \end{aligned} \tag{3}$$

Furthermore, we can compute that:

$$\begin{aligned} \mathrm{E}(Y_i^*|G_i, Z_i) &= \beta_0 + \beta_E \mu_E + (\beta_G + \beta_{G\times E}\mu_E) G_i + \beta_Z Z_i \\ \mathrm{Var}(Y_i^*|G_i, Z_i) &= (\beta_{G\times E} G_i)^2 \sigma_E^2 + \beta_E^2 \sigma_E^2 + 1 \\ Y_i^*|G_i, Z_i &\sim \mathrm{N}\Big(\beta_0 + \beta_E \mu_E + (\beta_G + \beta_{G\times E}\mu_E) G_i + \beta_Z Z_i, (\beta_{G\times E} G_i)^2 \sigma_E^2 + \beta_E^2 \sigma_E^2 + 1\Big) \end{aligned} \tag{4}$$

That implies that the probability we get a case for different levels of G_i and Z_i will be:

$$\begin{aligned} \mathrm{P}(Y=1|G_i,Z_i) &= \mathrm{P}(Y^*>0|G_i,Z_i) \\ &= \mathrm{P}\left(\frac{Y^*-\mathrm{E}(Y^*|G_i,Z_i)}{\sqrt{\mathrm{Var}(Y^*|G_i,Z_i)}} > \frac{-\mathrm{E}(Y^*|G_i,Z_i)}{\sqrt{\mathrm{Var}(Y^*|G_i,Z_i)}}\right) \\ &= \Phi\left(\frac{\mathrm{E}(Y^*|G_i,Z_i)}{\sqrt{\mathrm{Var}(Y^*|G_i,Z_i)}}\right) \end{aligned} \tag{5}$$

Therefore, applying the inverse CDF on both sides, we get

$$\Phi^{-1}\Big(\mathrm{P}(Y=1|G,Z)\Big) = \frac{\beta_0+\beta_E\mu_E+(\beta_G+\beta_{G\times E}\mu_E)G_i+\beta_Z Z}{\sqrt{(\beta^2_{G\times E}G_i^2\sigma_E^2+\beta_E^2\sigma_E^2+1)}}$$

This is only a linear function of G_i when the interaction parameter $\beta_{G\times E}=0$, and the slope of Z is constant across different genes only when $\beta_{G\times E}$ is zero.

1. If the true underlying model also contains another regressor W but W is uncorrelated with G for example. Then even though ignoring that regressor breaks the structural assumption of probit model, so that the fitted model without W is no longer a probit model (since now ϵ does not follow standard normal), but $\Phi^{-1}(\mathrm{P}(Y_i=1|G_i,Z_i))$ will still be a linear function of G_i. So detecting based on the linearity of $\Phi^{-1}\mathrm{P}$ will not be affected by omitted exogenous regressors.
2. Since $P(Y_i=1|G_i,Z_i)$ is actually unknown in practice, we can estimate it using the sample proportion $\hat{P}(Y=1|G=g,Z=z)=\frac{\sum_{i=1}^n \mathrm{I}\{y_i=1,G_i=g,Z_i=z\}}{\sum_{i=1}^n \mathrm{I}\{G_i=g,Z_i=z\}}$. We shouldn't use the fitted model to estimate them since our fitted model may be wrong.
3. The reason we used probit model instead of logistic model here is that assuming E follows normal distribution, $Y^*|G,Z$ will still be normal if we omit the interaction term, since linear combination of normal is normal. But assuming E follows logistic distribution does not imply that $Y^*|G,Z$ will be logistically distributed as logistic distribution is not closed under linear combination. However, based on the literature, it seems like probit model and logistic model have really closed results in real applications.

If the model is Genotypic instead:

$$Y_i^* = \beta_0+\beta_{G1}I(G_i=1)+\beta_{G2}I(G_i=2)+\beta_Z Z_i+\beta_E E_i+\beta_{G1E}I(G_i=1)\times E_i+\beta_{G2E}I(G_i=2)\times E_i+\epsilon_i \tag{6}$$

then we can derive the following:

$$\Phi^{-1}\Big(\mathrm{P}(Y=1|G,Z)\Big) = \frac{\beta_0+\beta_E\mu_E+(\beta_{G1}+\beta_{G1E}\mu_E)I(G_i=1)+(\beta_{G2}+\beta_{G2E}\mu_E)I(G_i=2)+\beta_Z Z}{\sqrt{(\beta^2_{G1E}I(G_i=1)\sigma_E^2+\beta^2_{G2E}I(G_i=2)\sigma_E^2+\beta_E^2\sigma_E^2+1)}}$$

In this case, the model will still be linear in $I(G_i=1), I(G_i=2)$ because of the extra parameter, just with different regression parameters. But the slope of Z will continue to differ between different Genetic types unless the there are no interaction effects β_{G1E} and β_{G2E}.

2 Method for Additive Model:

In this section, I will present two methods for the detection of interaction effect when the true model is additive.

2.1 Testing of Linearity:

Recall that when the model is additive, then:

$$\Phi^{-1}\Big(\mathrm{P}(Y=1|G,Z)\Big) = \frac{\beta_0+\beta_E\mu_E+(\beta_G+\beta_{G\times E}\mu_E)G_i+\beta_Z Z}{\sqrt{(\beta^2_{G\times E}G_i^2\sigma_E^2+\beta_E^2\sigma_E^2+1)}}$$

This method relies on the checking of linearity of $\Phi^{-1}(P)$, so the test statistics will also be focusing on the detection of linearity. We first derive the test statistic using the Delta method on sample proportions, and then we will generalize this test statistic using a linear regression framework.

2.1.1 Derivation using Delta Method:

If the two competing models are:

$$\begin{aligned} 1.Y_i^* &= \beta_0+\beta_G G_i+\beta_E E_i+\epsilon_i \\ 2.Y_i^* &= \beta_0+\beta_G G_i+\beta_E E_i+\beta_{G\times E}G_i\times E_i+\epsilon_i \end{aligned} \tag{7}$$

Let $p_i = P(Y=1|G=i)$, which can be approximated by the sample proportion $\hat{p_i}$. We know that

$$\Phi^{-1}(\hat{p_i}) \sim N\Bigg(\Phi^{-1}(p_i), \frac{1}{\phi(\Phi^{-1}(p_i))^2}\frac{p_i(1-p_i)}{n_i}\Bigg)$$

Here n_i denotes the number of $G=i$ in our dataset. Let $v_i = \frac{1}{\phi(\Phi^{-1}(\hat{p_i}))^2}\frac{\hat{p_i}(1-\hat{p_i})}{n_i}$ denote the estimate of $Var(\Phi^{-1}(\hat{p_i}))$.

Let $S = \Phi^{-1}(\hat{p_2})-2\Phi^{-1}(\hat{p_1})+\Phi^{-1}(\hat{p_0})$, and $T = \frac{S^2}{v_o+4v_1+v_2}$ be our test statistic. If 1 is the true model, then $T \sim X_1^2$.

2.1.2 Two Stage regression framework:

The above relationship can be generalized using the idea of a two stage regression. In the first stage, we consider the following generalized linear regression:

$$E(Y_i|G_i) = \Phi\Big(\beta_0+\beta_1\mathrm{I}(G_i=1)+\beta_2\mathrm{I}(G_i=2)\Big)$$

From this generalized linear regression, we obtained the fitted linear predictors $\hat{\eta_i} = \hat{\beta}_0+\hat{\beta}_1\mathrm{I}(G_i=1)+\hat{\beta}_2\mathrm{I}(G_i=2)$. Then, we fit the following weighted least square regression in the second stage:

$$\hat{\eta_i} = \gamma_0+\gamma_1\mathrm{I}(G_i=1)+\gamma_2\mathrm{I}(G_i=2)+\epsilon_i$$

where $\epsilon_i \sim N\Big(0, \frac{\Phi(\hat{\eta_i})(1-\Phi(\hat{\eta_i}))}{\phi(\hat{\eta_i})^2}\Big)$. Define the weight matrix $W = \mathrm{diag}\{\frac{\Phi(\hat{\eta_i})(1-\Phi(\hat{\eta_i}))}{\phi(\hat{\eta_i})^2}\}$, then our previous test statistic is just the test statistic of Wald test of $H_0: \gamma_2 = 2\gamma_1$.

The intuition behind how this approach works to detect the interaction between G and E can be thought as following. In the first regression, we are actually trying to get an estimate of $\eta = \frac{\mathrm{E}(Y^*|G)}{\sqrt{\mathrm{Var}(Y^*|G)}}$. Recall from the first section, if there is interaction between G and E, then η will be non-linear in G due to the fact that $\mathrm{Var}(Y^*|G)$ will be non-constant (We cannot directly test on whether Y^* has constant variance because it is not estimable). Therefore, if we observe that $\hat{\eta}$ is linear in G, which means $\gamma_2 = \gamma_1$, then we can be reasonable confident that there is no interaction present in the model.

We can use the first stage regression to obtain accurate estimate of η because regardless whether η is linear in G, we know η can be written as a linear model with $\mathrm{I}(G_i = 1)$ and $\mathrm{I}(G_i = 2)$. This suggests, if we have another covariate Z in the model, we can just do the two stage regressions with all of $\mathrm{I}(G_i = j)$ and $\mathrm{I}(G_i = j)Z_i$ (Then in our null hypothesis, besides that $\gamma_2 = 2\gamma_1$, we can also add additional constraints that all of $\mathrm{I}(G_i = j)Z_i$ should have the same slopes to increase our power).

2.2 Testing using random slopes:

First, let's rewrite our previous latent variable specification:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_G G_i + \beta_Z Z_i + \beta_{G\times E} G_i \times E_i + \epsilon_i \\ &= \beta_0 + \beta_G G_i + \beta_Z Z_i + U_i * G_i + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \\ U_i &= \beta_{G\times E} * E_i \end{aligned} \tag{8}$$

Here U_i can be thought as a random effect (random slope), being drawn from distribution $\mathrm{N}(0, \sigma_u^2)$. Notice that $\sigma_u^2 = \beta_{G\times E}^2 \sigma_E^2$. Therefore, testing for $\beta_{G\times E} = 0$ is equivalent to testing $\sigma_u^2 = 0$ for the random effects. In this case, we do not need to restrict our distribution to the probit model anymore. Since both probit model and logistic model are flexible enough to incorporate an observations-level random slopes. (There shouldn't be any identifiability problem with have too many random slopes(same number as observations), as including an observations-level random intercepts is a common trick to account for overdispersion in Poisson regression.)

In this case, we can use the likelihood Ratio test to test the model with and without the random slopes, with correction to the boundary. Therefore, the final test statistic will be $-2\Lambda \sim 0.5X_1^2 + 0.5X_0^2$ under null hypothesis.

2.2.1 Difference between two potential methods

1. The first method relies on the assumption that the true underlying model is probit model, and the distribution of E is normal. These assumptions shouldn't be too restrictive as it is said in the literature that probit model and logistic model tend to give similar results. However, the second method can be used for both probit model and logistic model. The only assumption in the second method is that E follows a normal distribution.

2. The next step for the first method is to develop a test statistic for testing the linearity. While for the second method, it seems like there are plenty of tools of testing at boundary to test $\sigma_u = 0$, using likelihood ratio. It seems like in the second method, jointly testing for the main effect and interaction effect

3. For the simulations of sample size 300000, the first method is very efficient to compute as it basically just computes nine sample proportions and compute their difference. If we can find a good test statistic for this, the hypothesis testing will be efficient to carry out and scale to larger sample. The second method takes a very long time to converge when the interaction is actually present in the model, and lme4 tends to give some warnings about the potential convergence problems if a probit model is fitted and underlying model has the interaction effect. For a larger sample with more regressors, the computational loads will be bigger for the second method.

4. Note that for the second method to work, we need to assume there is no main effect of the environment variable E_i, i.e. $\beta_E = 0$ need to be assumed in the model, because otherwise we will need not only individual level slopes but also individual level intercepts to capture the true model, which will cause identifiability issues.

3 Method for Genotypic Model:

3.1 Auxiliary variable method:

Recall for a Genotypic Model with interaction like below:

$$Y_i^* = \beta_0 + \beta_{G1} I(G_i = 1) + \beta_{G2} I(G_i = 2) + \beta_Z Z_i + \beta_E E_i + \beta_{G1E} I(G_i = 1) \times E_i + \beta_{G2E} I(G_i = 2) \times E_i + \epsilon_i \tag{9}$$

We can derive that:

$$\begin{aligned} \Phi^{-1}\Big(\mathrm{P}(Y = 1|G, Z)\Big) &= \frac{\beta_0 + \beta_E \mu_E + (\beta_{G1} + \beta_{G1E}\mu_E) I(G = 1) + (\beta_{G2} + \beta_{G2E}\mu_E) I(G = 2) + \beta_Z Z}{\sqrt{(\beta_{G1E}^2 I(G = 1)\sigma_E^2 + \beta_{G2E}^2 I(G = 2)\sigma_E^2 + \beta_E^2 \sigma_E^2 + 1)}} \\ &= \gamma_0 + \gamma_1 I(G = 1) + \gamma_2 I(G = 2) + \gamma_Z Z + \gamma_{Z1G} I(G = 1) \times Z + \gamma_{Z2G} I(G = 2) \times Z \end{aligned} \tag{10}$$

where the new parameters γ_{Z1G} and γ_{Z2G} will be defined as:

$$\gamma_{Z1G} = \frac{\beta_Z}{\sqrt{\beta_E^2 \sigma_E^2 + \beta_{G1E}^2 \sigma_E^2 + 1}} - \frac{\beta_Z}{\sqrt{\beta_E^2 \sigma_E^2 + 1}}$$

and:

$$\gamma_{Z2G} = \frac{\beta_Z}{\sqrt{\beta_E^2 \sigma_E^2 + \beta_{G2E}^2 \sigma_E^2 + 1}} - \frac{\beta_Z}{\sqrt{\beta_E^2 \sigma_E^2 + 1}}$$

In other words, a Genotypic Model with an missing interaction can still be written as a linear function of these two indicator functions of G because of the extra regression parameter. However, ignoring this environment to gene interaction will create an artificial interaction between gene and the covariate Z. Since the interaction effects are zero if and only if the covariate Z has constant slopes across different genotypes, we can test the environmental interaction by testing the null hypothesis $H_0 : \gamma_{Z1G} = \gamma_{Z2G} = 0$, using either wald test, likelihood ratio test or score test.

The key in this method is to test the equal slopes of the auxiliary variable Z. In order for this method to work, we need the following assumption:

1. The auxiliary variable Z_i is assumed to have no interaction effect with G in the model conditional on G, Z and E.
2. The auxiliary variable Z_i is also assumed to have no interaction effect with E in the model conditional on G, Z and E.