Progress report: Detecting interaction with unknown environmental covariate

Ziang Zhang

15/10/2020

1 Summary of current progress:

1.1 Latent Model for binary data

For binary response variable, it is often assumed that the reponse 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.

1.2 Method 1: Detection from linearity

1.2.1 When the true model does not contain interaction with environmental factor

First, consider that the true underlying model for the response variable Y_i is a probit model without interaction effect, i.e:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_G G_i + \beta_Z Z_i + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \\ \epsilon_i &\sim \mathrm{N}(0,1) \end{aligned} \tag{2}$$

Therefore, it can be shown that:

$$\begin{aligned} \mathrm{P}(Y_i = 1|G_i, Z_i) &= \mathrm{P}(\epsilon_i > -(\beta_0 + \beta_G G_i + \beta_Z Z_i)) \\ &= 1 - \Phi(-(\beta_0 + \beta_G G_i + \beta_Z Z_i)) \\ &= \Phi(\beta_0 + \beta_G G_i + \beta_Z Z_i) \end{aligned} \tag{3}$$

Where $\Phi(.)$ denote the CDF function of standard normal distribution. Therefore, $\Phi^{-1}\Big(\mathrm{P}(Y_i = 1|G_i, Z_i)\Big)$ shoud be a linear function of both G_i and Z_i.

1.2.2 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:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_G G_i + \beta_Z Z_i + \beta_{G\times E} G_i \times E_i + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \\ \epsilon_i &\sim \mathrm{N}(0,1) \end{aligned} \tag{4}$$

Furthermore, we can compute that:

$$\begin{aligned} \mathrm{E}(Y_i^*|G_i, Z_i) &= \beta_0 + (\beta_G + \beta_{G\times E}\mu_E)G_i + \beta_Z Z_i \\ \mathrm{Var}(Y_i^*|G_i, Z_i) &= (\beta_{G\times E} G_i)^2 \sigma_E^2 + 1 \\ Y_i^*|G_i, Z_i &\sim \mathrm{N}\Big(\beta_0 + (\beta_G + \beta_{G\times E}\mu_E)G_i + \beta_Z Z_i, \big(\beta_{G\times E} G_i\big)^2 \sigma_E^2 + 1\Big) \end{aligned} \tag{5}$$

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

$$\begin{aligned} \mathrm{P}(Y = 1|G_i, Z_i) &= \mathrm{P}(Y^* > 0|G_i, Z_i) \\ &= \mathrm{P}\Big(\frac{Y^* - \mathrm{E}(Y^*|G_i, Z_i)}{\sqrt{\mathrm{Var}(Y^*|G_i, Z_i)}} > \frac{-\mathrm{E}(Y^*|G_i, Z_i)}{\sqrt{\mathrm{Var}(Y^*|G_i, Z_i)}}\Big) \\ &= \Phi\Big(\frac{\mathrm{E}(Y^*|G_i, Z_i)}{\sqrt{\mathrm{Var}(Y^*|G_i, Z_i)}}\Big) \end{aligned} \tag{6}$$

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

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

This is not a linear function of G_i, but is a linear function of Z_i.

1. If the true underlying model also contains another regressor W but W is uncorrelated with G for example. Then eventhough 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.

1.2.3 Test Statistic of this method:

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. Depending on the type of data available, there will be several slightly different test statistics for different cases.

1.2.3.1 Case 1: When G is the only regressor in the model

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

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

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 $T \sim X_1^2$.

1.2.3.2 Case 2: When both G and Z are in the regression, where Z is discrete:

If the two competing models are:

$$\begin{aligned} 1.Y_i^* &= \beta_0 + \beta_G G_i + \beta_Z Z_i + \beta_E E_i + \epsilon_i \\ 2.Y_i^* &= \beta_0 + \beta_G G_i + \beta_Z Z_i + \beta_E E_i + \beta_{G\times E} G_i \times E_i + \epsilon_i \end{aligned} \tag{8}$$

Let $\hat{p_{ij}}$ denote the sample proportion of cases in the group with $G = i$ and $Z = j$, then we know that $\hat{p_{ij}}$ will be independent across different i and j. Also, by CLT:

$$\hat{p_{ij}} \sim N(p_{ij}, \frac{p_{ij}(1-p_{ij})}{n_{ij}})$$

where n_{ij} denote the number of observations in the (i,j) cell.

By delta method: we can obtain the distribution of $\Phi^{-1}(\hat{p_{ij}})$ being:

$$\Phi^{-1}(\hat{p_{ij}}) \sim N\left(\Phi^{-1}(p_{ij}), \frac{1}{\phi(\Phi^{-1}(p_{ij}))^2}\frac{p_{ij}(1-p_{ij})}{n_{ij}}\right)$$

where ϕ denotes the density of a standard normal.

Let $W_{ij} = \Phi^{-1}(\hat{p_{ij}})$. The variance of W_{ij} can be estimated as $v_{ij} = \frac{1}{\phi(\Phi^{-1}(\hat{p_{ij}}))^2}\frac{\hat{p_{ij}}(1-\hat{p_{ij}})}{n_{ij}}$, which is simply plugging $\hat{p_{ij}}$ for the unknown true probability p_{ij}. Let $S_1 = a_0(W_{10} - W_{00}) + a_1(W_{11} - W_{01}) + a_2(W_{12} - W_{02})$ and $S_2 = a_0(W_{20} - W_{10}) + a_1(W_{21} - W_{11}) + a_2(W_{22} - W_{12})$, where a_i is weight given to each difference term, such that $\sum_{i=0}^{2} a_i = 0$. If the frequency of G or Z is known. Then $a_i = P(Z = i)$ when we are testing for the interaction of G with E. So S_1 and S_2 will have a nice interpretation being estimated expected effect of G.

Under the null hypothesis that $\beta_{G\times E} = 0$ which means no interaction between G and E, we know that $\Phi^{-1}(p_{ij})$ should be linear in i. That is: $W_{(i+1)j} - W_{ij} \sim N(b_i, v_{(i+1)j} + v_{ij})$ for all j = 0,1,2. So:

$$S_1 \sim N\left(\sum_{i=0}^{2} a_i b_i, \sum_{i=0}^{2} a_i^2(v_{1i} + v_{0i})\right)$$

$$S_2 \sim N\left(\sum_{i=0}^{2} a_i b_i, \sum_{i=0}^{2} a_i^2(v_{2i} + v_{1i})\right)$$

with the covariance between S_1 and S_2 be denoted as C, which can be computed as:

$$C = \text{Cov}(S_1, S_2) = -\sum_{i=0}^{2} v_{1i} a_i^2$$

That means, if the null hypothesis is true,

$$T = \frac{(S_1 - S_2)^2}{\sigma_{S_1}^2 + \sigma_{S_2}^2 - 2C} \sim X_{df=1}^2$$

We will reject the null hypothesis when T has a large value.

Question to study: If G and Z are uncorrelated, we can just ignore the Z variable and use the test statistic in the first case. Which way will be better?

1.2.3.3 Case 3: When both G and Z are in the regression, where Z is continuous: Under this case, if G and Z are uncorrelated, we can just ignore the variable Z in the regression and use the test statistic for case 1.

However if G and Z are correlated, simply ignoring Z and use the test statistic in case 1 will produce invalid result (If significant p-value is obtained, it may be due to interaction with E, or due to correlation with Z).

To handle the case when Z is a continuous random variable correlated with G, we could use a likelihood based method, which involves fitting two models and compare the likelihood ratio. We are going to talk about that test statistic in the later section for joint testing of main effect and interaction effect of G.

1.2.4 Relationship with linear regression:

1.2.4.1 Regression with three points The test statistic proposed above can be viewed as the Wald test statistic of the following regression:

$$\begin{aligned}
&\Phi^{-1}(\hat{p_i}) = \beta_0 + \beta_1 \text{I}\{G_i = 1\} + \beta_2 \text{I}\{G_i = 2\} + \epsilon_i, \quad \text{for} \quad i = 0, 1, 2 \\
&\text{where} \quad \epsilon = (\epsilon_0, \epsilon_1, \epsilon_2) \sim \text{N}(0, \Omega) \\
&\text{where} \quad \Omega = \begin{bmatrix} v_0 & 0 & 0 \\ 0 & v_1 & 0 \\ 0 & 0 & v_2 \end{bmatrix}
\end{aligned} \tag{9}$$

This is a weighted least square with weight matrix Ω capturing the heterokedasticity. We are interested in testing $H_0 : 2\beta_1 - \beta_2 = L\beta = 0$ where $L = (0, 2, -1)$. The Wald statistics to test it will be:

$$\frac{(L\hat{\beta})^2}{L(X^T \Omega^{-1} X)^{-1} L^T}$$

This turns out to be exactly the same test statistic we used previously $\frac{S^2}{v_o + 4v_1 + v_2}$.

1.2.4.2 Two stage regression approach: 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:

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

From this generlized linear regression, we obtained the fitted linear predictors $\hat{\eta_i} = \hat{\beta_0} + \hat{\beta_1} \text{I}(G_i = 1) + \hat{\beta_2} \text{I}(G_i = 2)$. Then, we fit the following weighted least square regression in the second stage:

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

where $\epsilon_i \sim N\left(0, \frac{\Phi(\hat{\eta_i})(1-\Phi(\hat{\eta_i}))}{\phi(\hat{\eta_i})^2}\right)$. Define the weight matrix $W = \text{diag}\{\frac{\Phi(\hat{\eta_i})(1-\Phi(\hat{\eta_i}))}{\phi(\hat{\eta_i})^2}\}$, 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 $\eta = \frac{\text{E}(Y^*|G)}{\sqrt{\text{Var}(Y^*|G)}}$. 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 $\text{Var}(Y^*|G)$ 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 $\text{I}(G_i = 1)$ and $\text{I}(G_i = 2)$. This suggests, if we have another covariate Z in the model, we can just do the two stage regressions with all of $\text{I}(G_i = j)$ and $\text{I}(G_i = j)Z_i$ (Then in our null hypothesis, besides that $\gamma_2 = 2\gamma_1$, we can also add additional constraints that all of $\text{I}(G_i = j)Z_i$ should have the same slopes to increase our power).

1.2.5 Joint Testing of main effect and interaction effect of G:

Suppose that we have the following true latent model:

$$Y_i^* = \beta_0 + \beta_G G_i + \beta_Z Z_i + \beta_E E_i + \beta_{G\times E} G_i \times E_i + \epsilon_i$$

. Then like we have derived in the previous section, we know that:

$$Y_i^*|G_i, Z_i \sim \text{N}\left(\beta_0 + (\beta_G + \beta_{G\times E}\mu_E)G_i + \beta_Z Z_i, \left(\beta_{G\times E} G_i\right)^2 \sigma_E^2 + 1\right)$$

Assume without loss of generality that E has been standardized to $\text{N}(0, 1)$, and define $\gamma = \beta_{G\times E}^2$, then the above model can be simplified into:

$$Y_i^*|G_i, Z_i \sim \text{N}\left(\beta_0 + \beta_G G_i + \beta_Z Z_i,\ \gamma G^2 + 1\right)$$

This model is identifiable, and we can fit it using maximum likelihood estimation (Actually there is a Rpackage that can fit Y^* with variance $\exp(\gamma G)$, called heteroskedastic probit model in econometrics. However, for our proposed model, it seems like we do need to write a Rfunction by ourselves to get the MLE). We can do joint testing of interaction effect by testing $H_0 : \gamma = 0$, using likelihood ratio test. Since $\gamma \geq 0$, we do need to do a correction for boundary so our LRT test statistic follows $X_0^2/2 + X_1^2/2$ under null hypothesis.

If we want to jointly test $H_0 : \gamma = 0$ and $\beta_G = 0$, we can use a similar LRT procedure with null distribution being $X_1^2/2 + X_2^2/2$.

1.3 Method 2: By modeling the interaction term as a random slope:

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

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_G G_i + \beta_Z Z_i + \beta_{G\times E} G_i \times E_i + \epsilon_i \\ &= \beta_0 + \beta_G G_i + \beta_Z Z_i + U_i * G_i + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \\ U_i &= \beta_{G\times E} * E_i \end{aligned} \tag{10}$$

Here U_i can be thought as a random effect (random slope), being drawn from distribution $\text{N}(0, \sigma_u^2)$. 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.)

1.3.1 Test statistic for Method 2:

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.

Question to study: The test statistic for method 2 seems to be similar to the LRT test statistic used in method 1 for joint testing of main effects and interaction effects. Are they actually mathematically the same?

1.4 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 boudnary 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 efficent 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.

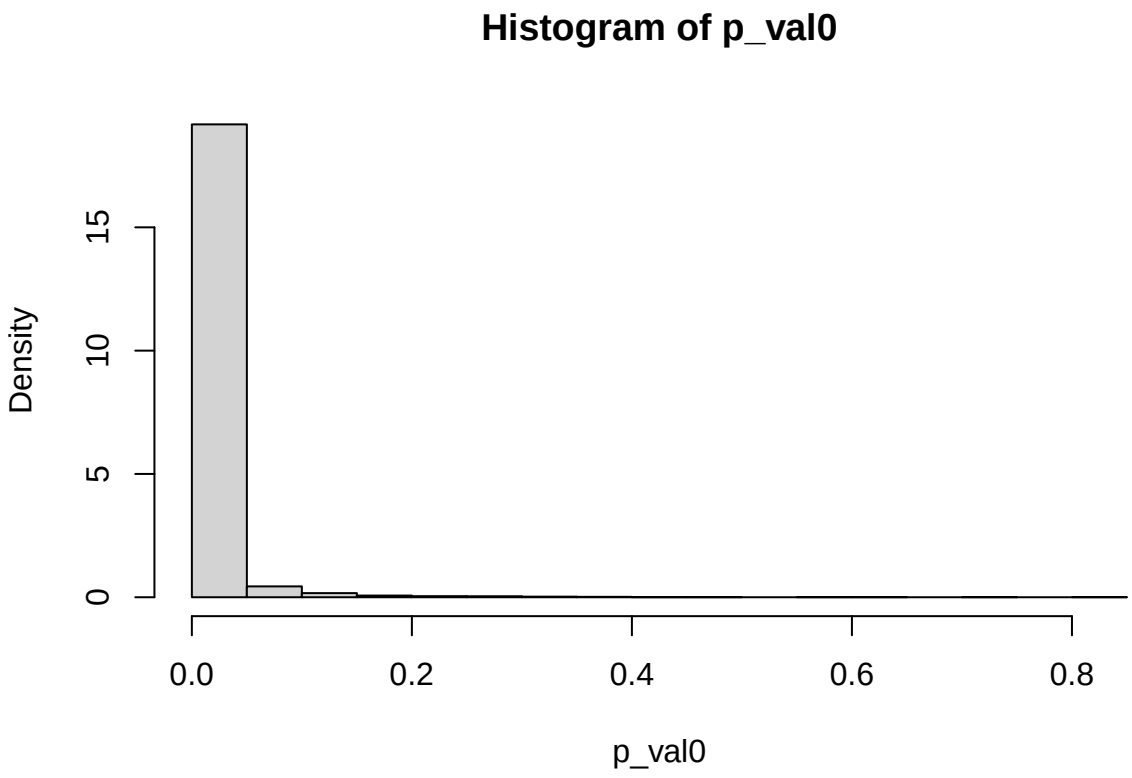
1.5 Simulation study:

Here we will implement our previously proposed methods, and see how they perform to detect the interaction

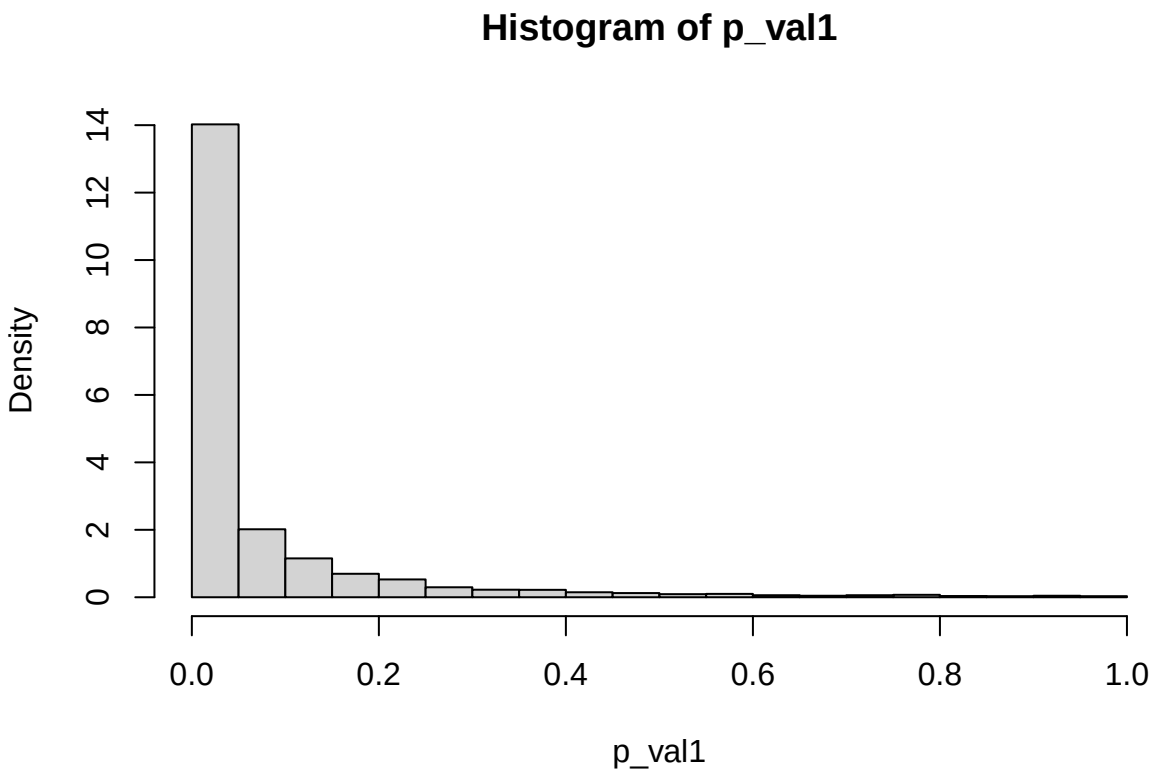
1.5.1 Method 1:

Take sample size being 3000, and assume that G has minor allele frequency 0.2 and E's distribution has been standardized. Let's do four examples with strong interaction effect, medium interaction effect, small interaction effect and negligible interaction effect:

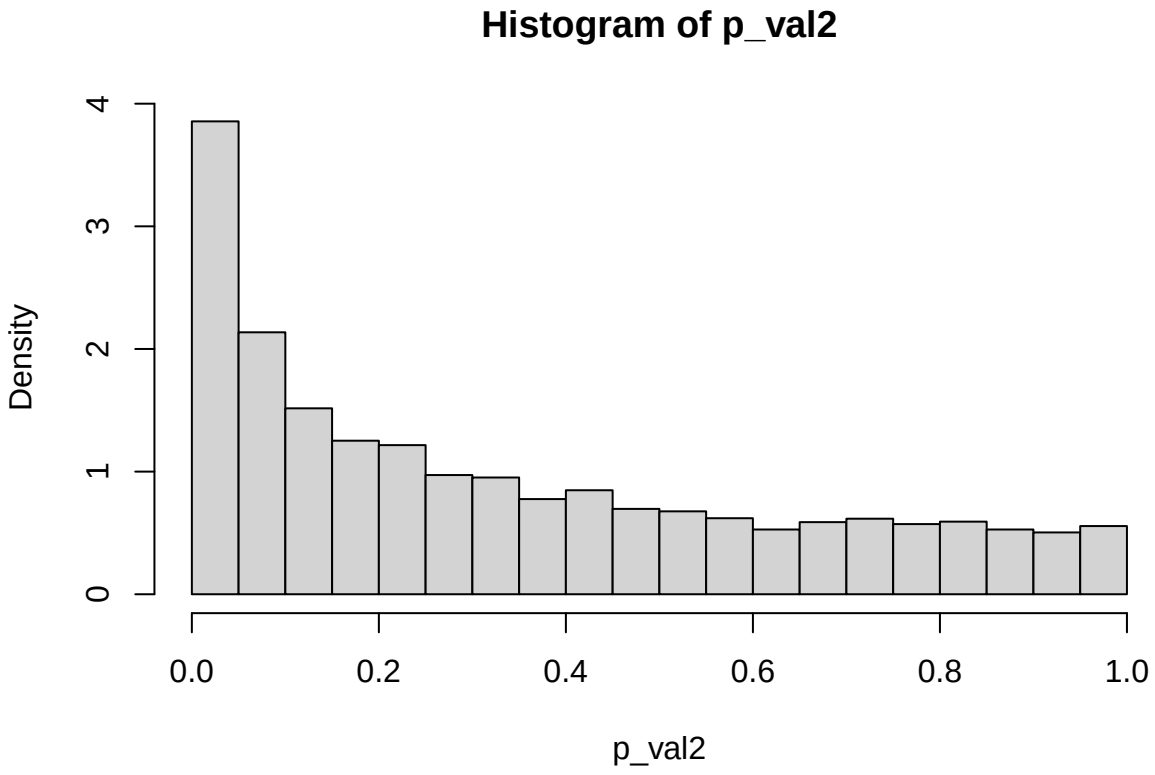
```
#### With strong interaction:
set.seed(123)
p_val0 <- Simulator_One_G(beta0 = -1.2, betaG = 0.8, betaE = 0.6, betaGE = 1.5, muE = 0, sdE = 1, p1 = 
hist(p_val0,breaks = 20, freq = F)
```



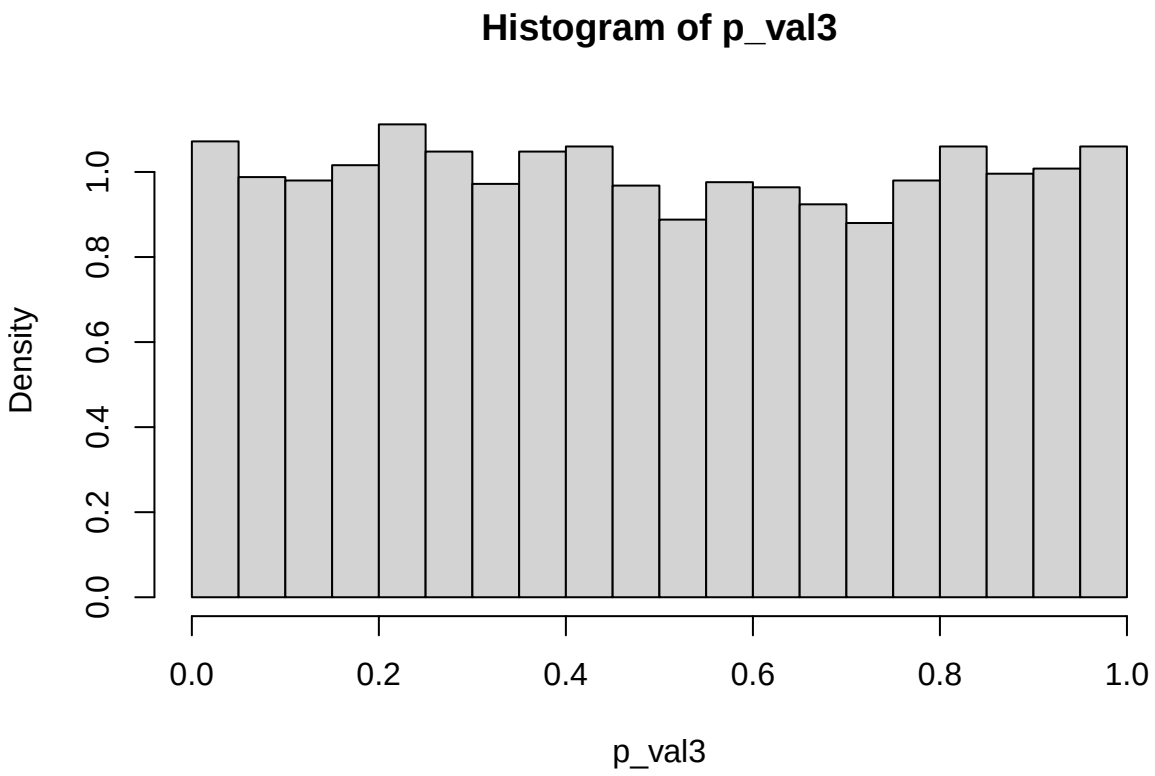

```
#### With medium interaction:
set.seed(123)
p_val1 <- Simulator_One_G(beta0 = -1.2, betaG = 0.8, betaE = 0.6, betaGE = 0.8, muE = 0, sdE = 1, p1 = 
hist(p_val1,breaks = 20, freq = F)
```




```
#### With small interaction:
set.seed(123)
p_val2 <- Simulator_One_G(beta0 = -1.2, betaG = 0.8, betaE = 0.6, betaGE = 0.3, muE = 0, sdE = 1, p1 = 
hist(p_val2,breaks = 20, freq = F)
```




```
#### With negligible interaction:
set.seed(123)
p_val3 <- Simulator_One_G(beta0 = -1.2, betaG = 0.8, betaE = 0.6, betaGE = 0.03, muE = 0, sdE = 1, p1 =
hist(p_val3,breaks = 20, freq = F)
```



1.5.2 Method 2:

This method takes much longer than the previous one, so we will not aggregate the p-values as before. The approach will be illustrated through two example: one without interaction and one with interaction:

```
set.seed(123)
n = 3000
p1 <- 0.2
q1 <- 1 - p1
beta0 <- -1.2
betaG <- 0.8
betaE <- 0.6
betaGE <- 0.6


### Without interaction:
G = apply(X = rmultinom(n,1,prob = c(p1^2,2*p1*q1,q1^2)) > 0, FUN = "which",MARGIN = 2) - 1
E <- rnorm(n, mean = 0, sd = 1)
latent_y <- beta0 + betaG*G + betaE*E  + rnorm(n = n)
y <- ifelse(latent_y > 0,1,0)
data <- data.frame(y = y, G = G)
data$OLRE <- 1:nrow(data)

model11 <- glmer(y~ G + (-1+G|OLRE), family = binomial(link = "probit"), data = data, nAGQ = 25L)
model12 <- glm(y~ G,family = binomial(link = "probit"), data = data)
```

```r
as.numeric((1 - pchisq(2*(logLik(model11) - logLik(model12)), df = 1))/2)
```

```
## [1] 0.1817477
```

```r
### With interaction:
set.seed(123)
latent_y <- beta0 + betaG*G + betaE*E + betaGE*G*E + rnorm(n = n)
y <- ifelse(latent_y > 0,1,0)
data <- data.frame(y = y, G = G)
data$OLRE <- 1:nrow(data)
model21 <- glmer(y~ G + (-1+G|OLRE) ,family = binomial(link = "probit"), data = data, nAGQ = 25L)
model22 <- glm(y~ G, family = binomial(link = "probit"), data = data)
as.numeric((1 - pchisq(2*(logLik(model21) - logLik(model22)), df = 1))/2)
```

```
## [1] 1.201479e-06
```