Detecting interaction with unknown environmental covariate

Ziang Zhang

15/10/2020

1 Summary of current idea:

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_1, G_2 come from a latent model such that:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \end{aligned} \tag{1}$$

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.

1.2 Potential Method 1: By checking 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:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \\ \epsilon_i &\sim \mathrm{N}(0,1) \end{aligned} \tag{2}$$

Therefore, it can be shown that:

$$\begin{aligned} \mathrm{P}(Y_i = 1|G_1, G_2) &= \mathrm{P}(\epsilon_i > -(\beta_0 + \beta_1 G_1 + \beta_2 G_2)) \\ &= 1 - \Phi(-(\beta_0 + \beta_1 G_1 + \beta_2 G_2)) \\ &= \Phi(\beta_0 + \beta_1 G_1 + \beta_2 G_2) \end{aligned} \tag{3}$$

Where $\Phi(.)$ denote the CDF function of standard normal distribution. Therefore, $\Phi^{-1}\Big(\mathrm{P}(Y_i = 1|G_1, G_2)\Big)$ shoud be a linear function of both G_1 and G_2.

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:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_1 \times E + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \\ \epsilon_i &\sim \mathrm{N}(0,1) \end{aligned} \tag{4}$$

Furthermore, we can compute that:

$$\begin{aligned} \mathrm{E}(Y_i^*|G_1,G_2) &= \beta_0 + (\beta_1 + \beta_3\mu_E)G_1 + \beta_2 G_2 \\ \mathrm{Var}(Y_i^*|G_1,G_2) &= (\beta_3 G_1)^2\sigma_E^2 + 1 \\ Y_i^*|G_1,G_2 &\sim \mathrm{N}\Big(\beta_0 + (\beta_1 + \beta_3\mu_E)G_1 + \beta_2 G_2, \big(\beta_3 G_1\big)^2\sigma_E^2 + 1\Big) \end{aligned} \tag{5}$$

That implies that the probability we get a case for different levels of G_1 and G_2 will be:

$$\begin{aligned} \mathrm{P}(Y=1|G_1,G_2) &= \mathrm{P}(Y^*>0|G_1,G_2) \\ &= \mathrm{P}\Big(\frac{Y^* - \mathrm{E}(Y^*|G_1,G_2)}{\sqrt{\mathrm{Var}(Y^*|G_1,G_2)}} > \frac{-\mathrm{E}(Y^*|G_1,G_2)}{\sqrt{\mathrm{Var}(Y^*|G_1,G_2)}}\Big) \\ &= \Phi\Big(\frac{\mathrm{E}(Y^*|G_1,G_2)}{\sqrt{\mathrm{Var}(Y^*|G_1,G_2)}}\Big) \end{aligned} \tag{6}$$

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

$$\Phi^{-1}\Big(\mathrm{P}(Y=1|G_1,G_2)\Big) = \frac{\beta_0 + (\beta_1 + \beta_3\mu_E)G_1 + \beta_2 G_2}{\sqrt{(\beta_3^2 G_1^2\sigma_E^2 + 1)}}$$

.

This is not a linear function of G_1, but is a linear function of G_2.

1. If the true underlying model also contains another regressor Z but Z is uncorrelated with G_2 for example. Then eventhough ignoring that regressor breaks the structural assumption of probit model, so that the fitted model without Z is no longer a probit model (since now ϵ does not follow standard normal), but $\Phi^{-1}(\mathrm{P}(Y_i=1|G_1,G_2))$ will still be a linear function of G_2. So detecting based on the linearity of $\Phi^{-1}\mathrm{P}$ will not be affected by omitted exogenous regressors.
2. Since $P(Y_i=1|G_1,G_2)$ is actually unknown in practice, we can estimate it using the sample proportion $\hat{P}(Y=1|G_1=g_1,G_2=g_2) = \frac{\sum_{i=1}^n \mathrm{I}\{y_i=1,G_{1i}=g_1,G_{2i}=g_2\}}{\sum_{i=1}^n \mathrm{I}\{G_{1i}=g_1,G_{2i}=g_2\}}$. 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, $Y^*|G_1,G_2$ 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 $Y^*|G_1,G_2$ will be logistically distributed as logistic distribution is not closed under linear combination. However, based on the literatures, it seems like probit model and logistic model have really closed results in real applications.
4. If this method is feasible, I will try to find a test statistic that has a nice asymptotic null distribution for the testing of linearity.

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

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

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_1 \times E_i + \epsilon_i \\ &= \beta_0 + \beta_1 G_1 + \beta_2 G_2 + U_i * G_1 + \epsilon_i \\ Y_i &= I\{Y_i^* > 0\} \\ U_i &= \beta_3 * E_i \end{aligned} \tag{7}$$

Here U_i can be thought as a random effect (random slope), being drawn from distribution $\mathrm{N}(0, \sigma_u^2)$. Notice that $\sigma_u^2 = \beta_3^2 \sigma_E^2$. Therefore, testing for $\beta_3 = 0$ 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.)

1.4 Test Statistic for Method 1:

Let $\hat{p_{ij}}$ denote the sample proportion of cases in the group with $G_1 = i$ and $G_2 = j$, 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 $Z_{ij} = \Phi^{-1}(\hat{p_{ij}})$. The variance of Z_{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 $S_1 = a_0(Z_{10} - Z_{00}) + a_1(Z_{11} - Z_{01}) + a_2(Z_{12} - Z_{02})$ and $S_2 = a_0(Z_{20} - Z_{10}) + a_1(Z_{21} - Z_{11}) + a_2(Z_{22} - Z_{12})$, where a_i is weight given to each difference term, such that $\sum_{i=0}^{2} a_i = 0$. If the allele frequency of G_1 or G_2 is known. Then $a_i = P(G_2 = i)$ when we are testing for the interaction of G_1 with E. So S_1 and S_2 will have a nice interpretation being estimated expected effect of G_1.

Under the null hypothesis that $\beta_3 = 0$ which means no interaction between G_1 and E, we know that $\Phi^{-1}(p_{ij})$ should be linear in i. That is: $Z_{(i+1)j} - Z_{ij} \sim N(b_i, v_{(i+1)j} + v_{ij})$ 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:

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

That means, if the null hypothesis is true,

$$T = \frac{(S_1 - S_2)^2}{\sigma^2_{S_1} + \sigma^2_{S_2} - 2C} \sim X^2_{df=1}$$

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

1.4.1 Simulation Study:

Let n = 3000, and allele frequencies for G_1 and G_2 being $p = 0.6$ and assuming that HWE holds in the population. First, consider the case when the true model does not have interaction between G_1 and E, let $\beta_0, \beta_1, \beta_2$ be $-1.2, 0.8, 0.3$ respectively.

```
set.seed(123)

n = 3000
p1 <- 0.6
q1 <- 0.4

p2 <- 0.6
q2 <- 0.4


##### Generate random genotype for G1 and G2, and a normal environmental factor that is unknown:
G1 = apply(X = rmultinom(n,1,prob = c(p1^2,2*p1*q1,q1^2)) > 0, FUN = "which",MARGIN = 2) - 1
G2 = apply(X = rmultinom(n,1,prob = c(p2^2,2*p2*q2,q2^2)) > 0, FUN = "which",MARGIN = 2) - 1


### Case 1: If the true model is nicely additive without interaction (Assuming probit model, inverse no
beta0 <- -1.2
beta1 <- 0.8
beta2 <- 0.3
latent_y <- beta0 + beta1*G1 + beta2*G2 + rnorm(n = n)
y <- ifelse(latent_y > 0,1,0)
```

We can do a test for G_1 first:

```
## Assuming that the true allele frequency of G2 is known
test_lin(y,G1,G2, weight = c(p2^2,2*p2*q2,q2^2))
```

```
## [1] 0.6514815
```

```
## If the true allele frequency of G2 is unknown, use equal weights:
test_lin(y,G1,G2)
```

```
## [1] 0.3209513
```

The p-values from our hypothesis testing are larger than 0.05, regardless whether we use the true allele frequency as the weights or not. We can do the same procedure to test G_2 as well, and the results should not be significant as well:

```
## Assuming that the true allele frequency of G2 is known
test_lin(y,G2,G1, weight = c(p1^2,2*p1*q1,q1^2))
```

```
## [1] 0.3457912
```

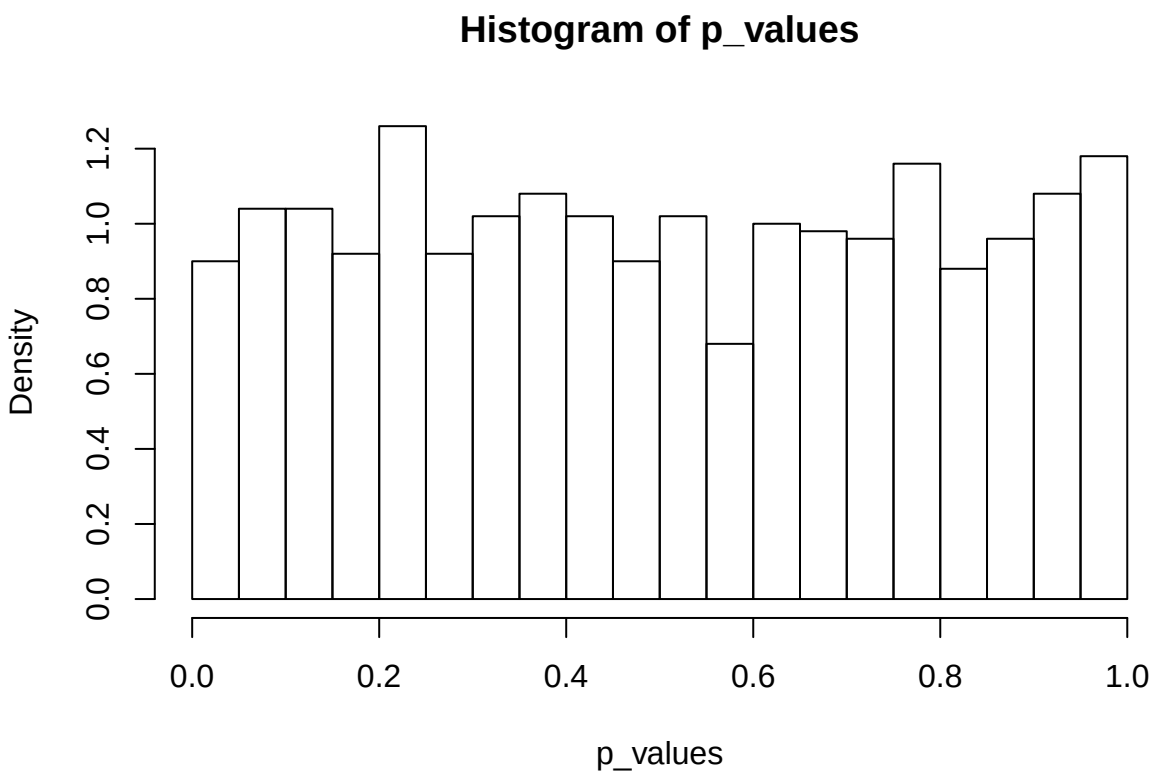
```
## If the true allele frequency of G2 is unknown, use equal weights:
test_lin(y,G2,G1)
```

```
## [1] 0.258314
```

The p-values also show that there should be no significant interaction between G_2 and E.

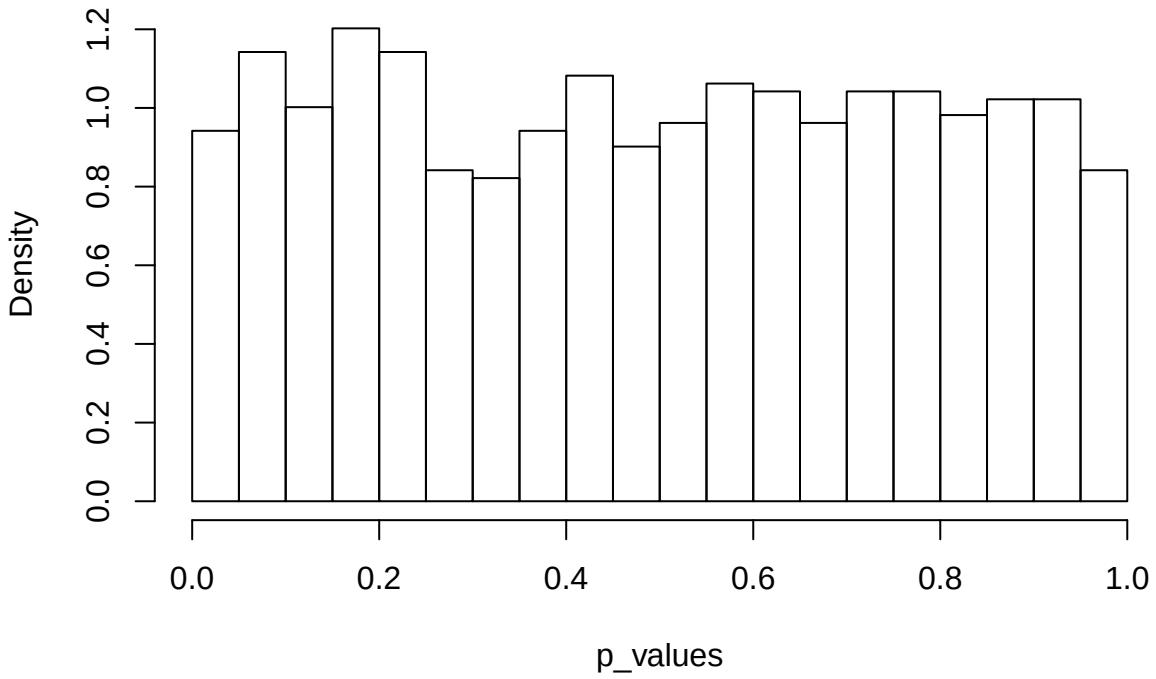
We can check the behavior of the p-values from our test by doing this simulation many times. Let's rerun the procedure for 1000 times and see how the p-values behave: first using the true allele frequency as weights:

```
p_values <- simulate_test(n = 3000, p1 = 0.6, p2 = 0.6, inter = FALSE, true_weight = T,num_trial = 1000
hist(p_values,freq = F,breaks = 30)
```



The histogram seems to be similar to uniform$[0, 1]$. Let's try again using equal weights:

```
p_values <- simulate_test(n = 3000, p1 = 0.6, p2 = 0.6, inter = FALSE, true_weight = F,num_trial = 1000
hist(p_values,freq = F,breaks = 30)
```



It still seems like uniform$[0, 1]$, so which weight to use won't affect much when the null hypothesis is true.

Then, let's see whether our test can detect that when there is a interaction $G_1 * E$ with $\beta_3 = 0.5$ in our model, where $E \sim N(1, 36)$.

```
set.seed(123)

n = 3000
p1 <- 0.6
q1 <- 0.4

p2 <- 0.6
q2 <- 0.4


##### Generate random genotype for G1 and G2, and a normal environmental factor that is unknown:
G1 = apply(X = rmultinom(n,1,prob = c(p1^2,2*p1*q1,q1^2)) > 0, FUN = "which",MARGIN = 2) - 1
G2 = apply(X = rmultinom(n,1,prob = c(p2^2,2*p2*q2,q2^2)) > 0, FUN = "which",MARGIN = 2) - 1
E <- rnorm(n, mean = 1, sd = 6)

### Case 2: If the true model does have interaction
beta0 <- -1.2
beta1 <- 0.8
beta2 <- 0.3
beta3 <- 0.5
latent_y <- beta0 + beta1*G1 + beta2*G2 + beta3*G1*E + rnorm(n = n)
y <- ifelse(latent_y > 0,1,0)

### using true frequency as weight:
test_lin(y,G1,G2, weight = c(p2^2,2*p2*q2,q2^2))
```

```
## [1] 0
```

```
### using equal weights:
test_lin(y,G1,G2)
```

```
## [1] 2.174927e-13
```

We can see that regardless whether we use the true frequency in our hypothesis testing, we have strong evidence to say that G_1 and E have interaction. Let's test in this case, whether G_2 has interaction with E:

```
test_lin(y,G2,G1, weight = c(p1^2,2*p1*q1,q1^2))
```

```
## [1] 0.1306443
```

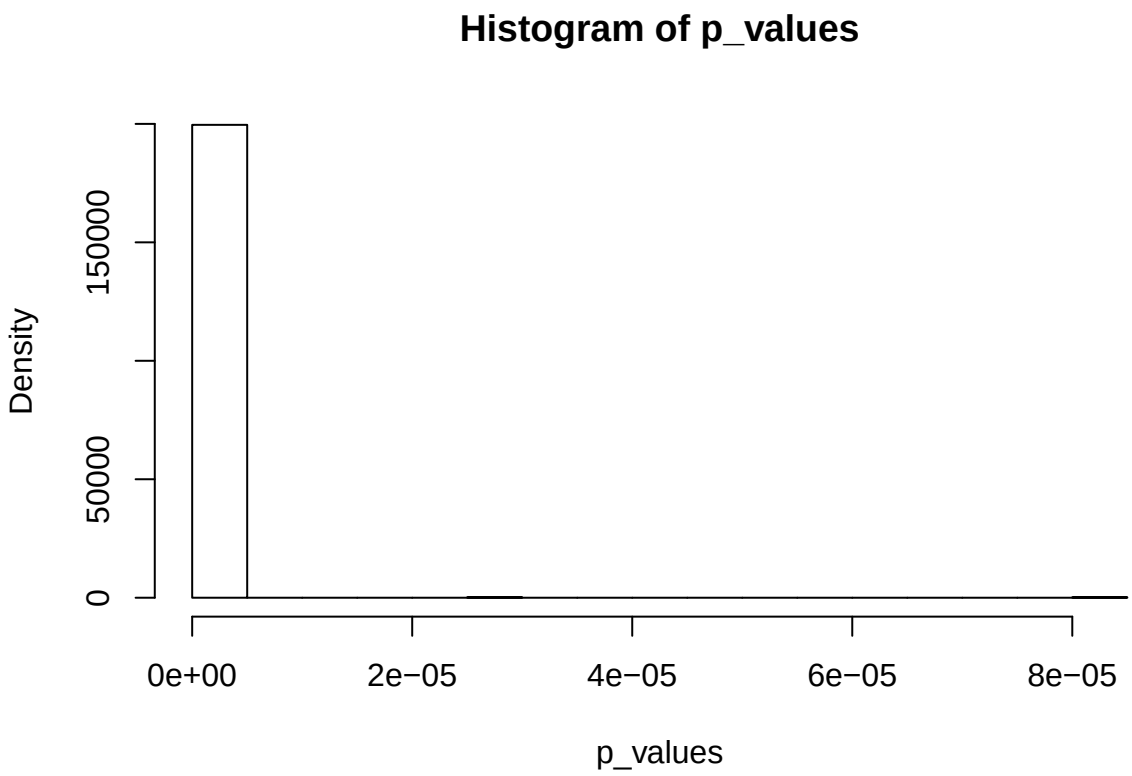
```
### using equal weights:
test_lin(y,G2,G1)
```

```
## [1] 0.4002292
```

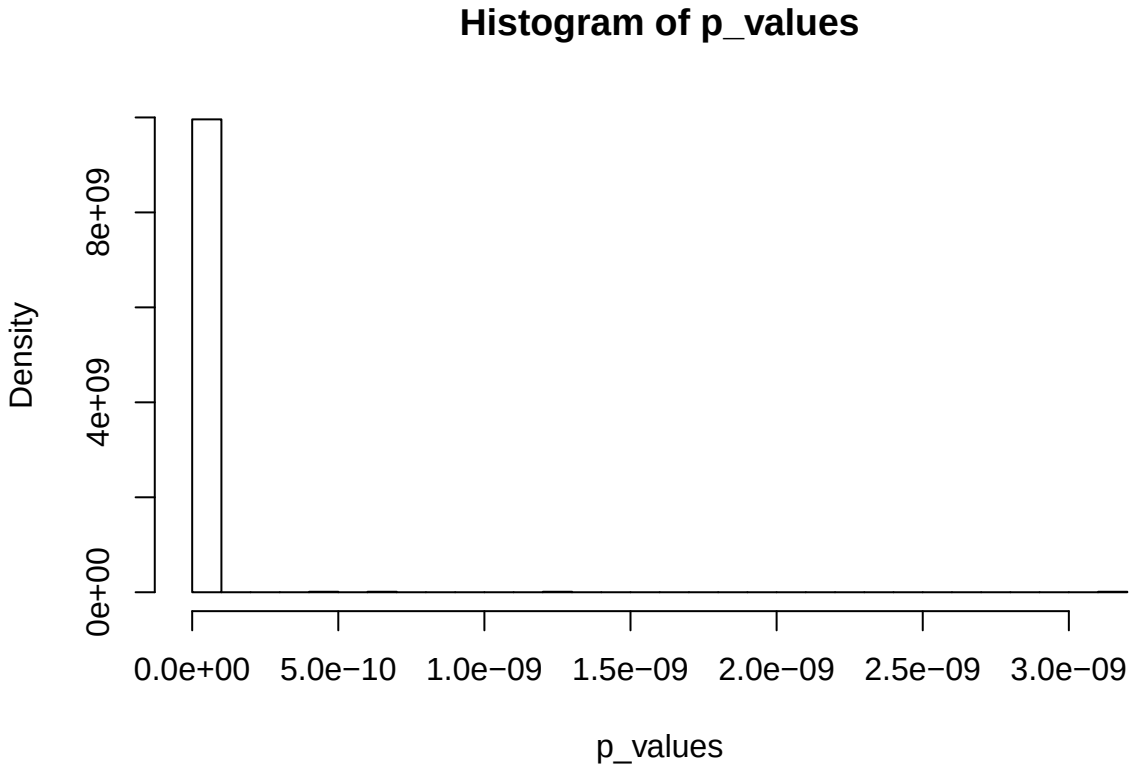
Like we expected, in this case, we still have evidence to say that G_2 does not have interaction with E.

Again, let's rerun this procedure 1000 times to see how powerful it is in this setting:

```
p_values <- simulate_test(n = 3000, p1 = 0.6, p2 = 0.6, inter = TRUE, true_weight = F,num_trial = 1000,
hist(p_values,freq = F,breaks = 30) ### Using equal weights
```




```
p_values <- simulate_test(n = 3000, p1 = 0.6, p2 = 0.6, inter = TRUE, true_weight = T,num_trial = 1000,
hist(p_values,freq = F,breaks = 30) ### Using true frequency
```



Regardless which weights being used, the distribution of p-values show that under this setting, the procedure has enough power to detect the interaction for all the 1000 simulations.

1.5 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.