Effect of measurement error on estimation of Gene-Environment interaction

Rima Rati Patra
MA23MSCST11015

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Guided by,
Dr. Arunabha Majumdar

Department of Mathematics,
Indian Institute of Technology, Hyderabad

Gene-Environment interaction

- Gene-environment interaction (GxE): The effect of genetics on traits or diseases is influenced by environmental factors.
- GxE is crucial for understanding variability in complex traits and diseases like diabetes, cancer, and heart conditions.
- For example, A genetic variant may increase disease risk, but the risk may only appear under specific environmental conditions such as diet, pollution, or lifestyle choices.

Statistical Model for GxE

To study GxE, statistical models evaluate whether the interaction term between genetic and environmental variables significantly contributes to the variation in the outcome.
Commonly used model is Multiple Linear Regression, given by:

$$y = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G * E$$

- Y is the phenotype.
- G, E and G*E are genotype, exposure and interaction respectively.
- β_0 is the intercept.
- β_1, β_2 are the main effect coefficients and β_3 is the interaction coefficient.

Measurement Errors

- Measurement error refers to the difference between the observed value of a variable and its true value.
- Both genetic and environmental factors are often subject to inaccuracies during measurement.
- These errors can distort the relationships between variables, undermining the validity of conclusions about gene-environment interactions (GxE).
- Errors can attenuate interaction effects, leading to underestimation of the true GxE effect and reduces the ability to detect significant GxE interactions.

Data Simulation

Data generative process: We have generated genotype and exposure data for 500 individuals.

- The genotype for each individual is simulated under Hardy-Weinberg equilibrium with the common allele (A) having a frequency P(A) =0.6.
- The probabilities of the resulting genotype are:

$$P(AA) = 0.36, P(Aa) = 0.48, P(aa) = 0.16.$$

- Genotypes are assigned numerical values:

$$AA = 2, Aa = 1, aa = 0$$

- Exposure is binary and the probability of getting exposed is 0.6 and drawn from Bernoulli Distribution

$$E \sim Bin(1, 0.6)$$

where E=1 refers to exposure and E=0 refers to no exposure.

Data generative process(Cont.)

- We set true parameters $\beta_0 \sim N(0,1), \beta_1 = 0.3, \beta_2 = 0.5, \beta_3 = 0.2$ and get the true phenotype Y as

$$Y = \beta_0 + 0.3G + 0.5E + 0.2G \times E$$

- We add noise to 10% of the genotype and 20% exposure data.
- After adding noise, we apply linear regression on both the true model (using true values for genotype and exposure) and the model with observed (noisy) variables.
- Our focus is on the interaction coefficient β_3.
- We repeat this process 500 times, and we see the summarized pictures and tables about the interaction coefficient.

Figure

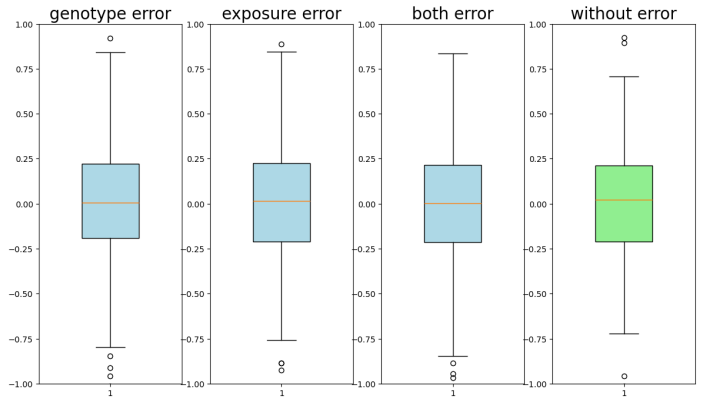

Figure: $\beta_3 = 0$

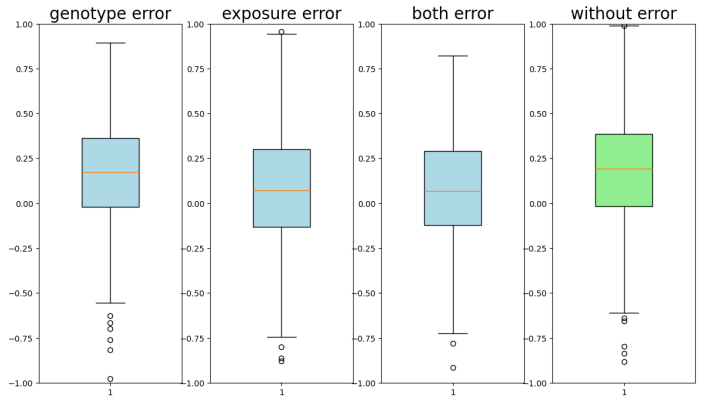

Figure: $\beta_3 = 0.2$

Figure interpretation

- **No Interaction:** The model estimates the interaction coefficient accurately even with measurement errors.
- **Presence of Interaction:** When the true interaction coefficient is 0.2, the model significantly underestimates it.
- **Effect of Errors:** The underestimation is more pronounced when both genotype and exposure errors are present, compared to when only the main effects have errors.

Bias and Error

Our true interaction coefficient β_3 is 0.2.
The impact of measurement errors was evaluated using the root mean squared error (RMSE) and relative bias of the estimated interaction coefficients $\hat{\beta}_3$ across multiple simulations.

- Relative bias is given as:

$$RelativeBias = \frac{\hat{\beta}_3 - \beta_3}{\beta_3} \times 100$$

- And Root Mean Square Error(RMSE) is given as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}((\hat{\beta}_3)_i - 0.2)^2}{N}}$$

Bias and Error

$\hat{\beta}_3$	Genotype error	Exposure error	Both errors	Without error
Mean	0.18	0.13	0.12	0.21
Relative bias (%)	-17	-56	-62	6
RMSE	0.007	0.015	0.018	0.004

Table: Summary of $\hat{\beta}_3$ under different error scenarios.

Observations

- Relative bias
  - Bias is lower with genotype error only compared to exposure error or both errors.
  - Errors in both genotype and exposure lead to larger bias in estimating the interaction.
- RMSE
  - RMSE is higher when both genotype and exposure errors are present.
  - This indicates greater distortion and reduced accuracy in estimation with both errors.

Hypothesis Testing and Statistical power

To evaluate the significance of the interaction term, we set up a hypothesis testing framework with the following:

- Null hypothesis H_0 : $\beta_3 = 0$
- Alternate hypothesis H_1 : $\beta_3 \neq 0$

For each scenario, hypothesis testing was performed on the β_3 using a two-tailed t-test The p-values were used to determine the significance of the interaction term at a 0.05 significance level. Here, the test statistic is

$$t^* = \frac{\hat{\beta}_3 - 0}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

where $\hat{\beta}$ is estimatesd coefficient , and σ is population standard deviation and n is population size. The p-value is calculated as the following probability

$$p - value = 2 * p(t \geq |t^*|) \tag{2}$$

Interpretation

Interpretation: Power = probability of rejecting null hypothesis when it is false

Scenarios	p-value	power
Genotype error	0.39	0.17
Exposure error	0.38	0.16
Both errors	0.4	0.14
WIthout error	0.37	0.18

- For $\beta_3 = 0.2$, the summarized p-value > 0.05(significance level). So in case of $\beta_3 = 0.2$, we fail to reject the null hypothesis, which is also indicated by the powers.
- The small effect size as 0.2 with errors in data, makes the test fail to detect the true effect.

Power Curve

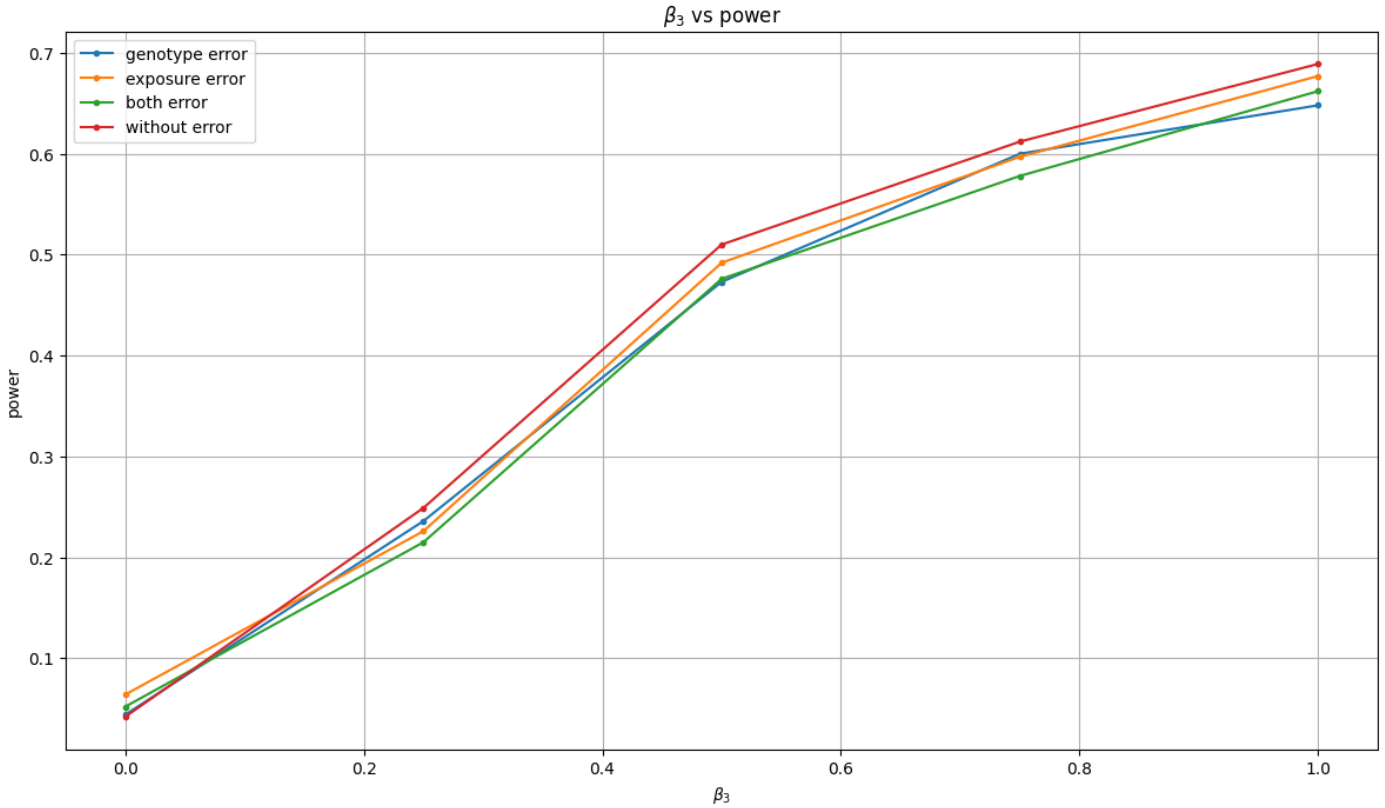

Figure: Power curve

Observation

- Across all scenarios, as β_3 increases, the power also increases. This suggests that the ability to detect the interaction effect improves with larger values of β_3.
- **Without Error:** Highest power throughout, indicating that data without errors yields the most accurate detection of the interaction .
- **Both Errors:** the lowest power at most points, reflecting the compounded impact of genotype and exposure errors on reducing detection ability.
- **Genotype vs. Exposure Error:** The power for genotype error is generally slightly higher than for exposure error, suggesting that exposure error may have a slightly more detrimental effect on power.

Thank You!